Imperial College London

Faculty of Medicine
Institute of Clinical Sciences
UKRI London Institute of Medical Sciences

Genetic effects of tissue-specific enhancers in schizophrenia and hypertrophic cardiomyopathy

Emanuele Felice Osimo

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London, 6th July 2023

Statement of originality

I hereby declare that the content of this thesis is the product of my own work carried out during my PhD and that all external sources have been properly referenced and acknowledged.

‘Education is an admirable thing, but it is well to remember from time to time that nothing that is worth knowing can be taught.’

Oscar Wilde, Intentions (1891)

Abstract

Most human conditions develop in genetically susceptible individuals from the interaction with environmental risk factors. These *complex* disorders result from the summation of effects from multiple genetic risk loci. Genome-wide association studies (GWASes) measure the association of single nucleotide polymorphisms (SNPs) with traits or conditions, and allow the creation of individualised polygenic risk scores. However, these explain only a small portion of a condition's genetic heritability. Further, there is evidence that schizophrenia GWAS signals are enriched within genomic regulatory blocks, which are clusters of conserved non-coding elements that span key developmental loci and function as long-range enhancers activating transcription of target developmental genes. This suggests that enhancer-based annotations might be useful to refine polygenic signals for schizophrenia.

In this work, I aimed to increase the amount of variance explained by PRS for schizophrenia, and a comparison condition hypertrophic cardiomyopathy, using tissue-specific regulatory enhancer-promoter annotations. To do so, I developed neural- and cardiac-specific enhancer lists, which I tested for enrichment, respectively, in schizophrenia and hypertrophic cardiomyopathy (HCM) heritability. I found that neural-specific enhancers are highly enriched in schizophrenia heritability – especially when overlapping genomic regulatory blocks. Then I created partitioned polygenic risk scores for enhancer-based and non-enhancer-based SNPs, where enhancer-based SNPs are prioritised. I further compared the amount of adjusted heritability for both conditions explained by original GWAS vs partitioned polygenic risk scores, and found up to a 6.5% increase in the Coefficient of Determination for schizophrenia, and similar amounts for HCM – however, this was not statistically significant. The increasing trend was specific for brain-expressed enhancers in schizophrenia, while it was widespread for HCM. Finally, I considered whether neural-specific enhancer-based partitions might be better modelled in GWAS using nonadditive effects, however my results were inconclusive due to small sample sizes.

Acknowledgements

A PhD is a lonely endeavour. For this reason, I would have struggled to get to the end of this PhD if it was not for several people, whom I would like to thank:

• From the Imperial Computational Regulatory Genomics lab, Boris, my supervisor, who has kindly kept me on a long leash and let me explore, but has also been there when I had technical or scientific questions.
Radina, a fellow student, who has been the humane face of the lab, and with whom I have shared the good and the hard times; she has also enabled this work by producing almost all of the data I have started from, as well as discussing how to best use the data for this work's benefit. Radina has also immensely helped my bioinformatics training by sharing code, and patiently reviewing my initial attempts. All other members of the lab, particularly Slava and Sarvesh, who have provided technical know-how and constructively discussed some of the aspects of this work.

• From the Psychiatric Imaging Group, Oliver, my co-supervisor, for all he has taught me. Oliver has scouted me in 2016, when I wanted to start in research and did not know how; he offered me a contract and enabled me to publish my first papers. He taught me a great deal: he's the person to go to when you have a research idea, to have it questioned and to be asked: "what is your specific hypothesis here?" many times, and always before starting the work. Oliver also taught me most of what I know about scientific writing, notably to be succinct and to never express concepts if they are not relevant and backed up by data.

• My external advisors, whose informal work has been completely thankless but almost as important as the formal supervision I had from Boris and Oliver. Graham Murray, at Cambridge, has been a great mentor, and has guided me multiple times when I was struggling to navigate my PhD. Graham also did a lot of work to connect me to our Cardiff collaborators, as well as discussing my findings at multiple time points, and suggesting where to go from there. Graham is another source of great learning and development on multiple fronts, including leadership and other transversal academic skills.
James Walters, at Cardiff, discussed my findings with us twice during my PhD, and he's intelligent insight into psychiatric genetics has been invaluable, as well as his practical ad-

vice. Andrew Morris at Manchester also provided some advice on partitioned PRSs and I am grateful for his time.

• My examiners, Paul Barton and Evangelos Vassos, who made the viva a pleasant and scientifically rewarding discussion.

• All the other collaborators to the present work here at Imperial, who have shared data and ideas, including James Ware and his group (cardiovascular genetics), including Sean and Francesco. Declan O'Regan's group (cardiovascular imaging and machine learning), who have collaborated and provided some ideas – including Antonio.

• My non-genetic collaborators, including Ben Perry, who is a friend and a colleague, and with whom I have endless scientific conversations on outcome prediction in psychosis; Toby Pillinger, who shares a passion with me and Ben for psychosis and its interactions with the body; Katherine Beck, who is a great friend and scientific mentor.

• My wife Francesca, my son Davide (who was born before the start of the PhD), and my son Camillo (who was born midway); once you have children, you put everything else into perspective – which helps when you are in an apparently endless tunnel – and I believe I have matured thanks to them. Having a family to "come back to" after work is one of those things that can never be overstated. Of course, I should also mention the very large number of sleepless nights, interrupted meetings and fewer opportunities to be at conferences...

• All the other people who have supported and helped me, most notably my wider family; Fabio – who left Imperial for Cambridge – as well as my friends.

Having nearly completed a PhD, it is still unclear to me if it represents a training and development opportunity, as its proponents declare, or if it is mostly a test of endurance. However, I have to say, a PhD is also a wonderful opportunity. For most of the time during my PhD, I have woken up every morning being grateful to be able to do what I love most, research, and even being paid for it. So, thank you to all taxpayers in the United Kingdom for generously funding me through the MRC and the NIHR.

Contents

Abstract **iii**

Acknowledgements **iv**

1 Introduction **1**

1.1 Genome structure and gene regulation by distal enhancers 1

1.1.1 The coding and non-coding genome 1

1.1.2 The 3D genome 4

1.1.2.1 Topologically Associated Domains or TADs 6

1.1.2.2 Detecting chromatin interactions 7

1.1.3 Enhancers, promoters and their interactions 8

1.1.3.1 Enhancer discovery with Cap Analysis of Gene Expression sequencing 9

1.1.3.2 Enhancer-promoter specificity 10

1.2 Extreme non-coding conservation 11

1.2.1 Genome conservation and conserved non-coding elements 11

1.2.2 Genomic regulatory blocks 12

1.3 The AR+C 14

1.3.1 Comparison of the AR+C with existing methods 16

1.4 Complex disorders and Genome-wide association studies (GWAS) 19

1.4.1 Studying complex human disease genetics 19

1.4.1.1 Complex disorders 20

1.4.1.2 Candidate genes and Linkage analyses 21

1.4.2 Genome-wide association studies (GWAS) 22

1.4.2.1 GWAS penetrance functions and models of inheritance . . . 22
1.4.3 Linkage disequilibrium . . . 25
1.4.3.1 LD and genotypes: tagging SNPs and clumping . . . 26
1.5 Polygenic risk scores and heritability of human disease . . . 28
1.5.1 Introduction to polygenic risk scores, or PRS . . . 28
1.5.1.1 Adjustment (shrinkage) of effect sizes . . . 29
1.5.1.2 Controlling for linkage disequilibrium and dealing with target populations . . . 30
1.5.1.3 PRS calculation . . . 30
1.5.2 Heritability and the missing heritability problem . . . 31
1.6 Introduction to schizophrenia and HCM . . . 35
1.6.1 Introduction to schizophrenia . . . 35
1.6.1.1 Schizophrenia genetics . . . 38
1.6.2 Introduction to HCM and its genetics . . . 41
1.7 Hypotheses and objectives . . . 43

2 Heritability enrichment in tissue-specific enhancers 47
2.1 Introduction . . . 47
2.1.1 Existing applications of the GRB model to psychiatric disorder genetics 47
2.1.2 Partitioned LD score regression (LDSC) . . . 48
2.1.3 Human enhancer annotations from the AR+C . . . 49
2.2 Materials . . . 50
2.2.1 Activation Ratio plus Contact (AR+C) . . . 50
2.2.1.1 Human-mouse GRBs . . . 52
2.2.2 PsychENCODE enhancers . . . 52
2.2.3 GTEx Tissue specific expression quantitative trait loci (eQTLs) . . . 53
2.3 Methods . . . 53
2.3.1 Generation of tissue-specific enhancer lists . . . 53
2.3.1.1 Tissue specificity of enhancer lists . . . 54
2.3.2 Measurement of the relative importance of enhancer-based partitions for specific GWAS using partitioned LDSC . . . 55

2.3.3 Software and code availability 56
2.4 Results 56
2.4.1 Tissue-specific enhancer partitions 56
2.4.1.1 Neural tissue and schizophrenia 57
2.4.1.2 Cardiac tissue and HCM 58
2.4.2 Tissue specific enhancers and heritability for schizophrenia and HCM 58
2.4.2.1 Neural-tissue-specific enhancers and schizophrenia heritability 61
2.4.2.2 Cardiac-tissue-specific enhancers and HCM heritability . . . 65
2.5 Summary of findings 69

3 Schizophrenia and HCM heritability from partitioned PRSs 70
3.1 Introduction 70
3.1.1 PRSs and risk prediction 70
3.1.2 Tissue-specific enhancers and associations with schizophrenia 73
3.2 Materials 75
3.2.1 Wave 3 Psychiatric Genomic Consortium Schizophrenia population and GWAS 75
3.2.2 Hypertrophic cardiomyopathy case-control GWAS and Royal Brompton Hospital target population 76
3.2.3 UK Biobank 76
3.3 Methods 77
3.3.1 PRS calculation 80
3.3.1.1 GWAS (base) and target quality control (QC) 80
3.3.1.2 GWAS (base) partitioning and clumping 81
3.3.1.3 Annotation and up-weighting of enhancer-based GWAS (base) partition SNP odds ratios 81
3.3.1.4 PRS calculation 82
3.3.2 Coefficient of determination calculation 83
3.3.3 Software 85
3.3.3.1 Code availability 85
3.4 Results 85

3.4.1 Neural tissue-specific enhancers and variance explained in schizophrenia 87
3.4.1.1 Patient and SNP selection for the *xs234* PGC schizophrenia target cohort . . . 87
3.4.1.2 Coefficients of determination for the main genomic partitions in schizophrenia . . . 89
3.4.1.3 Coefficients of determination for the original GWAS PRS vs multivariable models in schizophrenia . . . 92
3.4.1.4 Coefficients of determination for *enhancer-based* partitions in schizophrenia . . . 96
3.4.1.5 Can tissue-specific, *enhancer-based* partitioned PRSs help stratify people at risk for schizophrenia? . . . 99
3.4.2 Cardiac tissue-specific enhancers and variance explained in HCM . . . 102
3.4.2.1 Patient and SNP selection for the UK Biobank target cohort . 102
3.4.2.2 Coefficients of determination for the main genomic partitions in HCM . . . 103
3.4.2.3 Coefficients of determination for the original GWAS PRS vs multivariable models in HCM . . . 107
3.4.2.4 Coefficients of determination for *enhancer-based* partitions in HCM . . . 111
3.4.2.5 Can tissue-specific, *enhancer-based* partitioned PRSs help stratify people at risk for HCM? . . . 114
3.5 Summary of findings . . . 114

4 Leveraging nonadditive disease inheritance models 118
4.1 Introduction . . . 118
4.2 Methods . . . 120
4.2.1 Dominant and Recessive Schizophrenia EP-WAS in UK Biobank . . . 120
4.2.2 Internal and external validation . . . 121
4.2.3 Software and code availability . . . 122
4.3 Results . . . 122

4.3.1 Comparing Dominant and Recessive enhancer-based effects within UK Biobank . . . 123
4.3.2 EP-WAS internal validation in UK Biobank . . . 126
4.3.3 EP-WAS external validation in a PGC cohort . . . 129
4.4 Summary of findings . . . 130

5 Discussion **134**
5.1 Chapter 2: heritability enrichment in tissue-specific enhancers . . . 137
5.1.1 Interpretation of findings . . . 138
5.1.2 Limitations . . . 139
5.1.3 Further work . . . 140
5.2 Chapter 3: Schizophrenia and HCM heritability from partitioned PRSs . . . 141
5.2.1 Interpretation of findings . . . 143
5.2.2 Limitations . . . 145
5.2.3 Further work . . . 146
5.3 Chapter 4: Leveraging nonadditive disease inheritance models . . . 146
5.4 Conclusions . . . 149

Appendix A Supplementary Figures **151**
A.1 Sensitivity analyses: Chapter 3 results on celso and clz2a cohorts . . . 151
A.1.1 *clz2a* cohort . . . 151
A.1.2 *celso* cohort . . . 156
A.2 Sensitivity analyses: Chapter 3 results on xs234 – 0.05 threshold . . . 160
A.3 Sensitivity analyses: Chapter 3 results on HCM 0.05 . . . 165

References **169**

List of Figures

1.1 Organism complexity scales with the proportion of non-coding genomic elements. . . . 3
1.2 Hierarchical organisation of chromatin structure 5
1.3 The GRB model 13
1.4 Overview of the AR+C 15
1.5 AR+C benchmarking 17
1.6 A block of LD on human chromosome 9p21.3. 26
1.7 Heritability vs explained liability in select psychiatric conditions 33
1.8 The Clinical Course of Schizophrenia. 36
1.9 The Dopamine Hypothesis of Schizophrenia. 38
1.10 The normal vs hypertrophic heart. 42

2.1 LDSC for schizophrenia, main genomic partitions. 63
2.2 LDSC for schizophrenia, enhancer-based genomic partitions. 64
2.3 LDSC for HCM, main genomic partitions. 67
2.4 LDSC for HCM, enhancer-based genomic partitions. 68

3.1 PRS-related risk reduction for coronary heart disease and bladder cancer. . . 72
3.2 Chapter 3 graphical methods. 79
3.3 CoDs for the main partitions in schizophrenia. 90
3.4 CoDs for original vs multivariable models in schizophrenia - significant partitions. 93
3.5 CoDs for original vs multivariable models in schizophrenia - non-significant partitions. 94

3.6 CoDs for enhancer-based partitions in schizophrenia. 98
3.7 Double quantile plot for schizophrenia. 100
3.8 CoDs for the main partitions in HCM. 104
3.9 CoDs for the main partitions in HCM - RBH cohort. 106
3.10 CoDs for original vs multivariable models in HCM. 109
3.11 CoDs for original vs multivariable models in HCM - RBH. 110
3.12 CoDs for enhancer-based partitions in HCM. 112
3.13 CoDs for enhancer-based partitions in HCM - RBH. 113
3.14 Double quantile plot for HCM. 115

4.1 Schizophrenia ORs for enhancer-based SNPs – dominant model. 124
4.2 Schizophrenia ORs for enhancer-based SNPs – recessive model. 125
4.3 Schizophrenia EP-WAS UKBB validation: CoDs for the main partitions. . . . 127
4.4 Schizophrenia EP-WAS UKBB validation: CoDs for original vs multivariable models. 128
4.5 Schizophrenia EP-WAS PGC validation: CoDs for the main partitions. 131
4.6 Schizophrenia EP-WAS PGC validation: CoDs for original vs multivariable models. 132

A.1 Neural significant enhancers – clz2a. 152
A.2 Neural significant enhancers within GRBs – clz2a 153
A.3 Non neural enhancers – clz2a 154
A.4 Non associated enhancers – clz2a 155
A.5 Neural significant enhancers – celso 156
A.6 Neural significant enhancers within GRBs – celso 157
A.7 Non neural enhancers – celso 158
A.8 Non associated enhancers – celso 159
A.9 Neural significant enhancers – xs234 – 0.05 161
A.10 Neural significant enhancers within GRBs – xs234 – 0.05 162
A.11 Non neural enhancers – xs234 – 0.05 163
A.12 Non associated enhancers – xs234 – 0.05 164
A.13 CoDs for HCM at 0.05 threshold - Cardiac significant enhancers. 166

A.14 CoDs for HCM at 0.05 threshold - Non cardiac enhancers. 167
A.15 CoDs for HCM at 0.05 threshold - Non associated enhancers. 168

List of Tables

1.1 Distribution of cases and controls in a traditional GWAS at a single biallelc locus. 23

2.1 **Neural/brain-specific enhancer lists: size, enhancer-promoter distance, and tissue specificity** 59

2.2 **Cardiac-specific enhancer lists: size, enhancer-promoter distance, and tissue specificity** 60

Glossary

h^2	global disease heritability.
h^2_{SNP}	SNP-based heritability (usually from GWAS).
h^2_{pPRS}	partitioned PRS-derived heritability.
3C	Chromosome conformation capture.
3D	three dimensional.
AR+C	Activation Ratio plus Contact (AR+C) Method.
bp	base pair(s) (of DNA).
CAGE	cap analysis of gene expression.
cDNA	complementary or copy DNA; synthetic DNA that has been transcribed from a specific mRNA through a reaction using the enzyme reverse transcriptase.
CNE	conserved non-coding element.
CoD	coefficient of determination – or total variance explained by the genetic factor on the liability scale, corrected for ascertainment.
E-P	enhancer-promoter.
enh	enhancer.
EP-WAS	Enhancer-based GWAS.
eQTL	expression quantitative trait locus/loci.

eRNAs	enhancer-derived non-coding RNA.
ES	AR+C effect size = activation ratio.
FANTOM5	Atlas of mammalian promoters, enhancers, lncRNAs and miRNAs: https://fantom.gsc.riken.jp/5/.
FDR	false discovery rate.
GO	gene ontology.
GRB	genomic regulatory block.
GWAS	genome-wide association study.
HCM	hypertrophic cardiomyopathy.
Hi-C	Hi-C, a high-throughput Chromosome conformation capture method.
Kb(p)(s)	kilobase pair(s) (of DNA).
LD	linkage disequilibrium.
LDSC	LD score regression.
lncRNA	Long non-coding RNA.
LOO	leave-one-out.
MAF	minor allele frequency.
Mb(p)(s)	megabase pair(s) (of DNA).
mRNA	messenger RNA.
ncRNA	non-coding RNA.
PCR	polymerase chain reaction.
PGC	Psychiatric Genomic Consortium.
pPRS	*partitioned* polygenic risk score.

PRS	polygenic risk score.
QC	quality control.
SCZ	schizophrenia.
SNP	single nucleotide polymorphism.
TAD	topologically associating domain.
TF	transcription factor.
TPM	tags/transcripts per million.
TS	tissue specific.
TSS	transcription start site.
UKBB	UK Biobank.
UTRs	untranslated regions.

Chapter 1

Introduction

1.1 Genome structure and gene regulation by distal enhancers

1.1.1 The coding and non-coding genome

For many years, protein-coding genes have been the main subject of investigation in the human genome, despite the coding portions of genes, called exons, constituting merely 1.5% of the genome (Alexander et al., 2010). In the 1960s and early 1970s, when it was discovered that – in comparison for example with bacterial genomes – most of the human genome did not code for proteins, it was thought that these non-coding regions were non-functional, and named *junk DNA* (Ohno, 1972). The significance of the non-coding genome has gained increasing attention since more advanced sequencing efforts have uncovered the

true scale of the coding vs non-coding compartments (Alexander et al., 2010; Collins et al., 2003). Some started to notice how organism complexity seemed to scale more consistently with the amount of non-coding DNA over total, as compared to scaling with total DNA, as well exemplified by Taft et al., 2007 in Figure 1.1.

Subsequently, large-scale projects, such as the Encyclopedia of DNA Elements (ENCODE), have systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification across the human genome, making clear that non-coding elements in the genome far outnumber protein-coding genes, and starting to build evidence around the functional classes of non-coding elements, as we will see below (The ENCODE Project Consortium, 2012).

We now know that most of the genome in complex organisms is transcribed, and this transcription appears to vary over the course of development, which suggests that non-coding DNA transcripts have multiple roles across the development of complex organisms (Carninci et al., 2005). Non-coding regulatory DNA regions are now often found to act as transcription units, as exemplified by the widespread transcription observed at enhancers (Andersson et al., 2014; Djebali et al., 2012). Enhancers are genomic elements that function as distal regulatory sequences, capable of modulating spatio-temporal and quantitative gene transcription programs in response to environmental (external) or developmental (internal) stimuli (see paragraph 1.1.3 below).

In addition to genome regulatory regions such as enhancers, multiple classes of non-coding RNAs were discovered. Among the best studied non-coding RNAs (ncRNAs) are microRNAs (miRNAs), which can mediate post-transcriptional gene silencing by controlling the translation of mRNA into proteins. miRNAs are estimated to regulate the translation of more than 60% of protein-coding genes (Gebert & MacRae, 2019). miRNAs and other non-coding RNAs, such as small nucleolar RNAs (snoRNAs), PIWI-interacting RNAs (piRNAs), and the heterogeneous group of long non-coding RNAs (lncRNAs) have all been discovered to play a role in a number of medical conditions, including cancer, as well as neurological, cardiovascular, autoimmune, imprinting and monogenic disorders (Esteller,

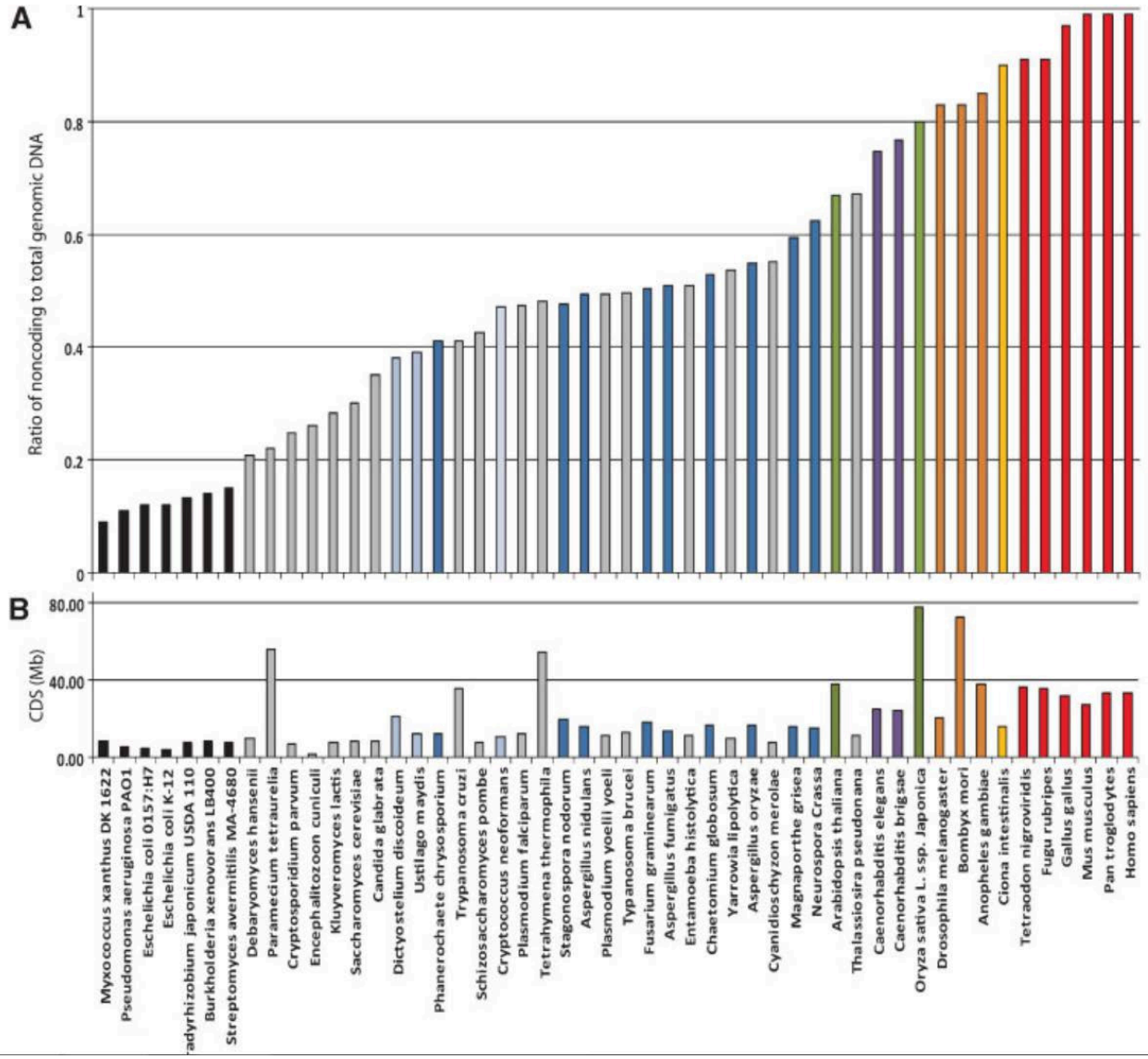


Figure 1.1: Organism complexity scales with the proportion of non-coding genomic elements. **Panel A** shows on the Y axis the fraction of non-coding DNA over total, while **Panel B** plots total genome size. For both panels, the X axis shows different organisms, sorted by organism complexity from left to right, from *Mixococcus* to *Homo Sapiens*.

2011).

Further, as we will see in section 1.4.1.1, increasing evidence points to the accumulation of loci associated with a variety of medical conditions within non-coding regions of the genome, and to the fact that common variation in non-coding regions could be at the heart of complex disorder aetiology (see section 1.4.2 and Maurano et al., 2012).

1.1.2 The 3D genome

The linear genome is highly compacted into the small nuclear space of each cell. To take humans as an example, to store over 2 metres of DNA (Piovesan et al., 2019) in the space of a cell nucleus, usually measuring a few µm in diameter, the DNA has to have a very organised 3D structure. While initially thought of primarily as a space-saving technique, over the last twenty years 3D folding has been shown to be functional, with studies highlighting the significance of spatial gene positioning for essential biological functions such as transcription, replication, and DNA repair among others (Therizols et al., 2014).

As shown in Figure 1.2, DNA folding is complex and multi-layered. The simplest level is that of nucleosomes, where DNA is wrapped around specific protein complexes called histones (Luger et al., 1997). At the kilobase-to-megabase scale (Figure 1.2a), **chromatin loops** form, which allow variable-range functional interactions between elements, such as enhancer-promoter interactions (Harmston & Lenhard, 2013). A major and relatively recent breakthrough was the discovery of stable regions characterised by high frequency of interactions between loci, which have been called **Topologically Associated Domains** or TADs (see section 1.1.2.1 below, and Figure 1.2b, (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). At the megabase scale, **chromatin compartments** are defined, that carry similar epigenetic marks and which usually encompass several TADs (Lieberman-Aiden et al., 2009). The formation of spatially segregated A and B compartments results in the separation of transcriptionally active chromatin regions (A compartments) from transcriptionally inactive, gene-poor regions (B compartments), see Figure 1.2c and Rao et al., 2014. At an even larger scale, chromatin is arranged into discrete **chromosome territories**, one for each chro-

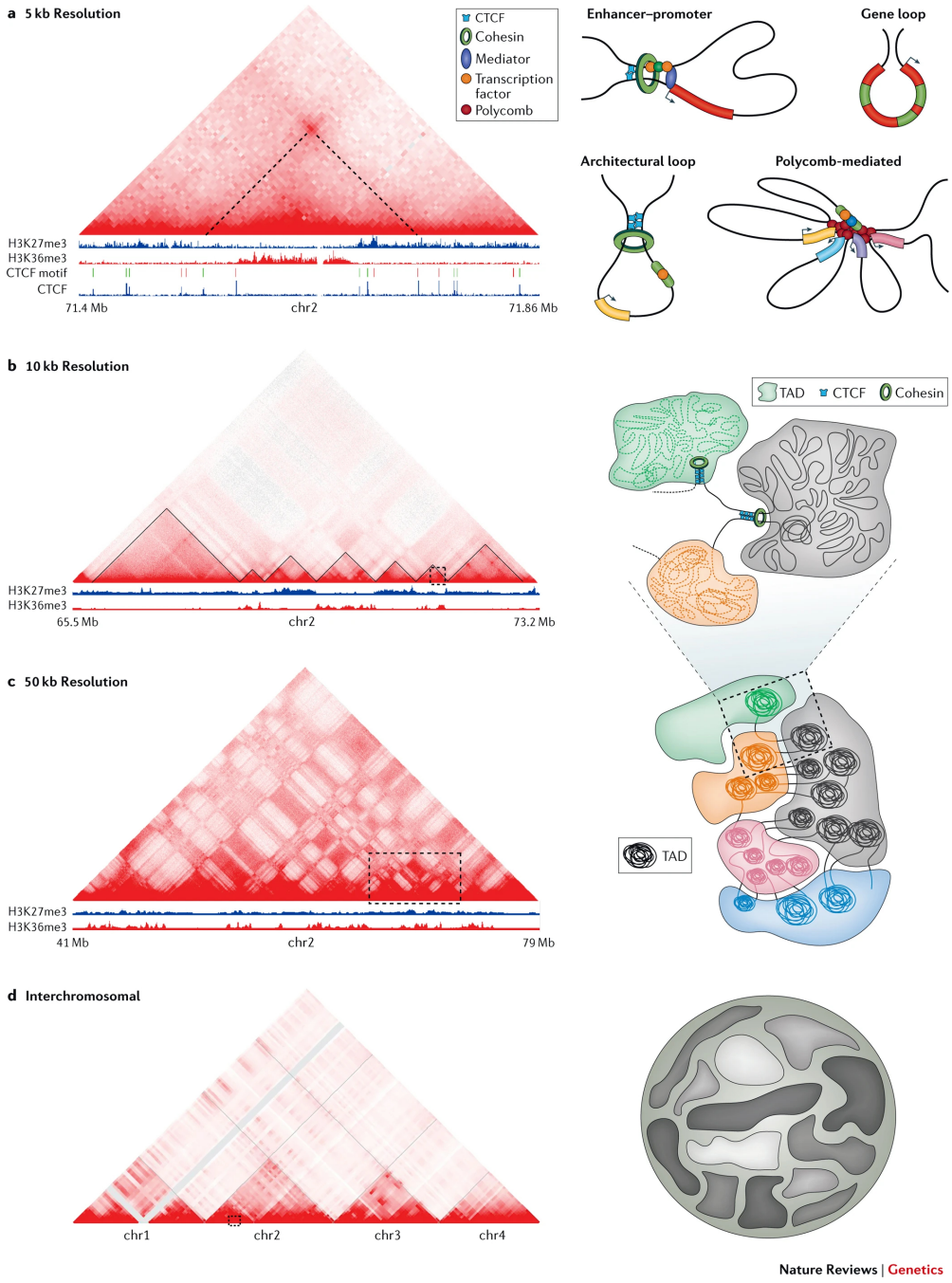


Figure 1.2: Hierarchical organisation of chromatin structure.
a) Types of **chromatin loops** that can reside within a domain (enhancer–promoter loop, Polycomb-mediated loop, gene loop or architectural loop). On the left is an example of an architectural loop as seen in high-resolution Hi-C data. b) Left: 8 Mb region containing several **TADs** as seen in Hi-C maps (TADs are manually annotated with solid lines). On the right, three different TADs, enriched for either active marks (H3K4me3 and H3K36me3; grey), Polycomb (H3K27me3; green) or heterochromatin (H3K9me3; orange) are schematically represented in the three-dimensional (3D) space. c) **Chromatin compartments** and d) **whole chromosome** scale, in both cases the Hi-C map is on the left, with a schematic representation of the 3D structure on the right.
Figure reproduced from Bonev and Cavalli, 2016; Licensed from Springer Nature through RightsLink License Number 5507640104734.

mosome. Hi-C experiments (see section 1.1.2.2) have shown that the interaction between these territories is infrequent (Figure 1.2d). As we will see below, each of these levels of folding appears today to be highly regulated, with some levels of folding appearing to be organism- and cell-specific, and others to be more widely conserved (Bonev & Cavalli, 2016).

1.1.2.1 Topologically Associated Domains or TADs

TADs are three-dimensional structures within the genome that consist of regions of DNA that are spatially close to each other, while being functionally separated from other regions of DNA (Pope et al., 2014). In the human genome, these regions are hundreds of kilobases to a few megabases in size, and are defined by a high frequency of chromatin interactions between them. TADs are formed by the looping of the chromatin fibre, which brings genes and regulatory elements into close proximity, enabling them to interact with each other, which is a prerequisite for gene regulation. TADs play a role in organising the genome into distinct functional units and in maintaining proper gene expression patterns during development and differentiation – as well as 'insulating', or separating, regions within a TAD from regions inside another one – thus preventing ectopic interactions between genes and regulatory elements (Beagan & Phillips-Cremins, 2020). Disruptions in TAD boundaries have been linked to a range of diseases and disorders, including malformation syndromes and developmental disorders (Flöttmann et al., 2015; Giorgio et al., 2015; Lupiáñez et al., 2015).

In conclusion, TADs are an important level of chromatin organisation, as they favour gene regulation by facilitating intra-TAD interactions, for example, between regulatory elements and genes (e.g., enhancers and promoters), even though these might be separated by large linear distances, while making inter-TAD interactions less likely; this is reviewed in Beagan and Phillips-Cremins, 2020.

1.1.2.2 Detecting chromatin interactions

Experimental assays are available to study how regions of DNA interact at a 3D level, the evolution and scale of which has allowed the enormous expansion of what we know today about 3D genome organisation. Initial discovery on this topic was slow and mostly limited to the chromosome scale, due to the need to directly visualise 3D conformations through the microscope (Stack et al., 1977). Subsequent innovations to the field of microscopy allowed gradual gains, including for example the advent of live-cell fluorescence microscopy, which has given insights into the dynamic properties of chromosome organisation (Heun et al., 2001).

However, it is the advent of Chromosome Conformation Capture, or **3C** techniques, that saw a step change in the scale of discovery (Dekker et al., 2002). 3C techniques involved cross-linking DNA and its associated proteins with formaldehyde, followed by digestion with a restriction enzyme to create DNA fragments that are then ligated together under conditions that favour the formation of intra- or inter-chromosomal ligation events. This produces chimeric DNA fragments that are then detected and quantified using PCR or sequencing. By analysing the frequency of ligation events between different regions of the genome, researchers can then infer the spatial proximity of these regions in the nucleus.

3C, however, can only capture interactions between two known loci at a time (De Wit & De Laat, 2012). More powerful variants of the 3C technique have been developed. One such technique is called **Hi-C**. Hi-C can capture genome-wide interactions at scale by producing high-throughput results, which has allowed to study chromatin compartmentalization, TADs and chromatin interactions genome-wide (Belton et al., 2012; Lieberman-Aiden et al., 2009). However, Hi-C is not suited for studying short-range regulatory interactions, which often occur at shorter distances (Lee et al., 2022). More recently, Hi-C has been refined to produce higher-resolution results, and a derived technique is called **Micro-C** (Hsieh et al., 2015); Micro-C uses Micrococcal nuclease (MNase) for enzymatic digestion, which produces shorter fragments, and therefore allows up to 1Kb resolutions in 3D chromosome mapping. See Figure 1.2 for examples of 3D genome interaction maps at different scales, ranging from

chromatin loops (top panel a) to chromosome scale (bottom panel d).

As we have seen, TADs are important structural features of 3D chromatin folding, which facilitate genetic interactions within, and hinder interactions outside of TAD boundaries. The importance of TADs for this work reside in the fact that TADs tend to overlap with GRBs, which we will introduce in section 1.2.2, and are structures favouring internal enhancer-promoter interactions, as we will see next in the next section.

1.1.3 Enhancers, promoters and their interactions

We have seen in section 1.1.1 that the majority of DNA is non-coding, and that the assessment of the importance of non-coding regions has gradually increased, as more and more of their structural and functional features are discovered. In this section, we are going to explore the important role that non-coding regulatory elements, and particularly enhancers, play in regulating gene expression, and why this is highly relevant to this work.

Two of the most important classes of gene expression regulatory elements are promoters and enhancers. Promoters span the transcription start site of genes, and are essential to initiate transcription (Haberle & Lenhard, 2016; Lenhard et al., 2012). Enhancers are *cis-non-coding regulatory elements* whose activity increases the expression of a target gene, i.e., they are positive regulators of gene expression. Their action can help coordinate changes in gene expression in space and time. Enhancer location relative to the gene promoter they control (target gene) varies from adjacent to the promoter, to many kilobases, and even megabases, upstream or downstream (e.g., in *cis*), and can even be located in the target gene's intron, or within other genes' introns (Schoenfelder & Fraser, 2019). Moreover, besides acting in a position-independent manner, enhancers can regulate transcription irrespective of their orientation (Banerji et al., 1983; Banerji et al., 1981; de Villiers & Schaffner, 1981; Moreau et al., 1981). Finally, one enhancer can regulate several genes, and at the same time each gene can be regulated by multiple enhancers (Osterwalder et al., 2018).

After the discovery of a number of enhancers (Banerji et al., 1981; Moreau et al., 1981), genomes were scanned more systematically for more by looking for elements with

the following characteristics: functional independence of the target gene promoter, hypersensitivity to DNase treatment (indicative of an accessible chromatin state), the presence of transcription factor binding sites, and enriched binding of transcription co-activators and histone acetylation (Bulger & Groudine, 2011). However, genome-wide annotation of enhancers based on these characteristics alone led to extremely large estimated number of enhancers in humans ($>$ 400,000 to $\sim$1 million), exceeding that of coding genes by more than ten-fold (Rivera & Ren, 2013).

1.1.3.1 Enhancer discovery with Cap Analysis of Gene Expression sequencing

A more recent approach to defining enhancers was linking them to their transcriptional activity. Two early studies showed that many of the sequences with the above-described epigenetic marks are transcribed into largely non-polyadenylated ncRNAs, which were named enhancer-derived ncRNAs or eRNAs (De Santa et al., 2010; Kim et al., 2010).

The later adoption of the **cap analysis gene expression (CAGE)** technique allowed enhancer annotation to proceed at pace. CAGE relies on *cap trapping* to capture 5′-complete complementary DNA fragments (cDNAs), reverse transcribed from 5′-capped mRNAs (Shiraki et al., 2003). The use of CAGE, paired with high-throughput techniques for sequencing, allowed to create an atlas of transcribed enhancers across different human and mouse tissues, and to detect $\sim$40,000–65,000 transcribed enhancers in humans (Andersson et al., 2014; Djebali et al., 2012). eRNA-producing enhancers were found to show higher binding of transcriptional co-activators, greater chromatin accessibility, and higher enrichment of active histone marks such as H3K27ac than those exclusively annotated by epigenomic marks (Andersson et al., 2014; Kim et al., 2010).

One of the early steps in enhancer activation is the binding of transcription factors (TFs). Despite some overlap in TF binding between enhancers and promoters, specific non-overlapping sets of TFs bind enhancers and promoters (The ENCODE Project Consortium, 2012). It is still not completely clear how different sets of TFs regulate enhancers and pro-

moters, and how exactly the interaction works, however a plausible model is that enhancers and promoters each independently recruit certain TFs, but require collaboration to achieve a full amplitude of transcriptional outputs (Hatzis & Talianidis, 2002).

As we have seen in section 1.1.2, genomes can compact and condense into complex 3D structures, which can bring together sections of DNA which would otherwise be distant. As we will see in the next section, the formation of enhancer-promoter loops has been shown to favour enhancer-promoter interactions (see section 1.1.2.1 and Schoenfelder and Fraser, 2019).

1.1.3.2 Enhancer-promoter specificity

Enhancer-promoter specificity refers to the capacity of enhancers to activate only their target genes and not other unrelated genes, which may be situated closer to them in the linear genomic sequence. While the molecular mechanisms and combinatorial logic responsible for this specificity are not yet well understood, several processes have been identified that contribute to the establishment of cell-type-specific transcriptional programs. Schoenfelder and Fraser, 2019 outlined these processes and their relationships in a three-step model, called the 'selecting-facilitating-specifying' model: the *selecting* step involves the binding of various factors that modify the chromatin state at cell-type-specific regulatory elements. The *facilitating* step concerns the folding of the chromatin fibre to promote spatial proximity between regulatory elements. Finally, the least understood step, called the *specifying* step, is thought to involve the stabilisation of specific enhancer-promoter interactions by proteins bound to those elements that preferentially interact with each other (Schoenfelder & Fraser, 2019).

In the next section we are going to see how non-coding elements can be classified on the basis of their phylogenetic conservation, and how this impacts function at a genomic level.

1.2 Extreme non-coding conservation

1.2.1 Genome conservation and conserved non-coding elements

More than thirty years ago, researchers first identified highly conserved sequences in the non-coding regions of metazoan genomes by comparing the introns and untranslated regions (UTRs) of mammalian and avian mRNAs. These studies discovered individual elements, without any apparent function, that had retained over 70% sequence identity for hundreds of millions of years of evolution (Hraba-Renevey & Kress, 1989; Yaffe et al., 1985).

These initial analyses were followed by genome-wide scans for similarly conserved regions – which identified hundreds to thousands of highly conserved non-coding elements that are traceable across more than 400 million years of evolution (Bejerano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005). The level of sequence conservation observed in these elements is often greater than that seen in protein-coding genes. These findings, together with the fact that these elements tend to cluster around genes encoding regulators of multi-cellular development and differentiation (Harmston et al., 2013; Sandelin et al., 2004), provided support to the functional significance of non-coding elements and suggested that they played crucial roles in the regulation of gene expression and the evolution of metazoan genomes (Bejerano et al., 2004; Engström et al., 2007; Kikuta et al., 2007b; Sandelin et al., 2004; Woolfe et al., 2005). These conserved sequences were collectively named *highly conserved noncoding elements* (HCNEs) (Kikuta et al., 2007a) or Conserved Non-Coding Elements (CNEs) (Polychronopoulos et al., 2017). Curated databases of CNEs exist for the benefit of researchers worldwide (Dimitrieva & Bucher, 2013; Engström et al., 2008).

The results of multiple studies utilising *in vivo* transgenic reporter assays across several animal species have led to the view that CNEs often act as enhancers, or cis-regulatory elements that help to coordinate the spatial-temporal expression of genes, and particularly during embryonic development (see for example De La Calle-Mustienes et al., 2005; Pennacchio et al., 2006; Visel et al., 2008; Woolfe et al., 2005). CNEs are thought to be important

components of the regulatory networks that govern gene expression and play a crucial role in the formation and maintenance of complex metazoan phenotypes (Sandelin et al., 2004; Sanges et al., 2013). In line with the view that CNEs are important in regulating animal development, diseases have been found to result from the mutation of CNE-resident enhancers: the most famous example of this is a mutations in the *sonic hedgehog SHH* ZRS enhancer, which results in preaxial polydactyly in both human and mouse (Lettice et al., 2003), even though many other examples exist (Becker & Rinkwitz, 2012; Miguel-Escalada et al., 2019; Navratilova & Becker, 2009; Ragvin et al., 2010).

1.2.2 Genomic regulatory blocks

CNEs are not randomly scattered across genomes: they are commonly found in clusters that span regions of low gene density, including gene deserts. These clusters, typically up to $\sim 2-5Mb$ long in humans, tend to include critical developmental regulatory target genes, and have been named Genomic Regulatory Blocks or **GRBs** (Bejerano et al., 2004; Engström et al., 2008; Kikuta et al., 2007b; Sandelin et al., 2004; Woolfe et al., 2005).

As represented in Figure 1.3, many CNEs within GRBs act as enhancers, promoting the transcription of a nearby **target gene** (Polychronopoulos et al., 2017). This can occur over large genomic distances, either across large gene deserts (Kikuta et al., 2007a; Nobrega et al., 2003), or in some cases by skipping over more proximal genes in the same locus, named **bystander genes** by Kikuta et al., 2007b, which remain unaffected by CNE-based regulation (Kleinjan & Van Heyningen, 2005; Lettice et al., 2003; Navratilova et al., 2009; Ragvin et al., 2010; Smemo et al., 2014).

As further shown in Figure 1.3, GRB boundaries tend to align with those of TADs in both vertebrates and invertebrates (Harmston et al., 2017). This suggests that GRB-associated TADs might possess distinct genomic characteristics. According to the GRB model, target and bystander genes have different expression patterns: most bystanders are broadly (ubiquitously) expressed, and typically correspond to housekeeping genes, while target genes tend to be developmentally regulated and more tissue-specific, although of-

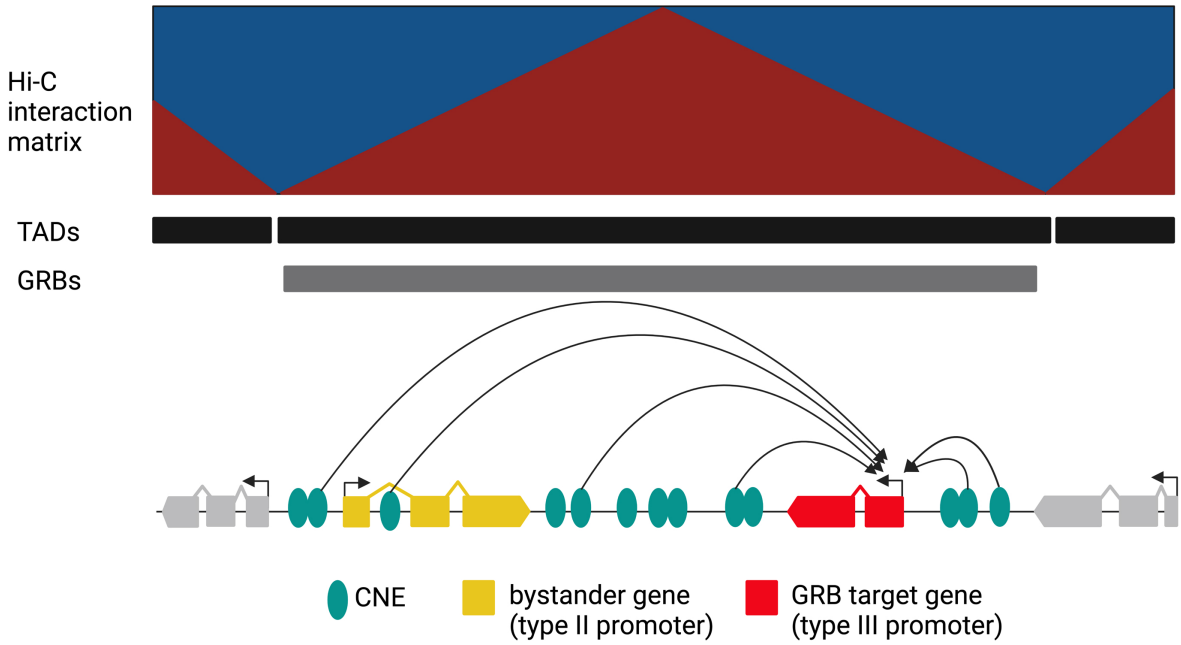


Figure 1.3: The GRB model.
The figure recapitulates the GRB model (bottom): multiple CNEs span a region, thus defining a GRB. Within this, most of these CNEs will act as enhancers, but the target gene will in most cases be one, with potentially multiple, unaffected, bystander genes.
The figure also shows that GRBs tend to overlap with TADs, shown as black blocks in the middle, which are shown to be highly self-interacting regions insulated from contact with other regions in the schematic Hi-C plot at the top.
The Figure was modified from Nash, 2018 by Georgieva, 2022. Reproduced here from with permission from the copyright holder.

ten expressed in multiple tissues, exhibiting complex spatiotemporal expression patterns (Akalin et al., 2009; Navratilova et al., 2009; Pennacchio et al., 2006). These findings support the notion that GRBs are involved in the precise regulation of gene expression during development.

1.3 Predicting target genes of long-range enhancers: introduction to the Activation Ratio plus Contact (AR+C) method

As we have seen in section 1.1.3.2, enhancers contribute to normal development and homeostasis by facilitating the precise regulation of their target genes in space and time. Dysregulation of this process can lead to disease, as demonstrated by the fact that:

- some conditions can directly stem from enhancer mutation or structural dysregulation involving disruptions of enhancer-promoter 3D interactions, e.g. brachydactyly in Lupiáñez et al., 2015 and activation of oncogenes in Flavahan et al., 2019;
- for complex disorders (see section 1.4.1.1), the majority of risk polymorphisms identified by genome-wide association studies (GWAS, see section 1.4.2) fall within non-coding regions of the genome, and a large proportion of these SNPs are predicted to overlap enhancers (Maurano et al., 2012).

The accurate identification of non-coding elements, including enhancers, and their associated target gene(s) is therefore crucial for studying the mechanisms underlying human disease. Although a common method for associating a non-coding element of interest to its target gene is to assign it to the nearest gene in the linear genomic sequence, this strategy ignores important aspects of spatial genome organisation and developmental gene regulation. Hence, such an approach may not provide an accurate understanding of the complex mechanisms that govern gene expression (Chua et al., 2022).

For all these reasons, a genome-wide dataset of regulatory enhancer-promoter (E-P) associations was generated by Georgieva, 2022, as part of her doctoral work in the Lenhard *Computational Regulatory Genomics* lab. This resource has been named the **Activation**

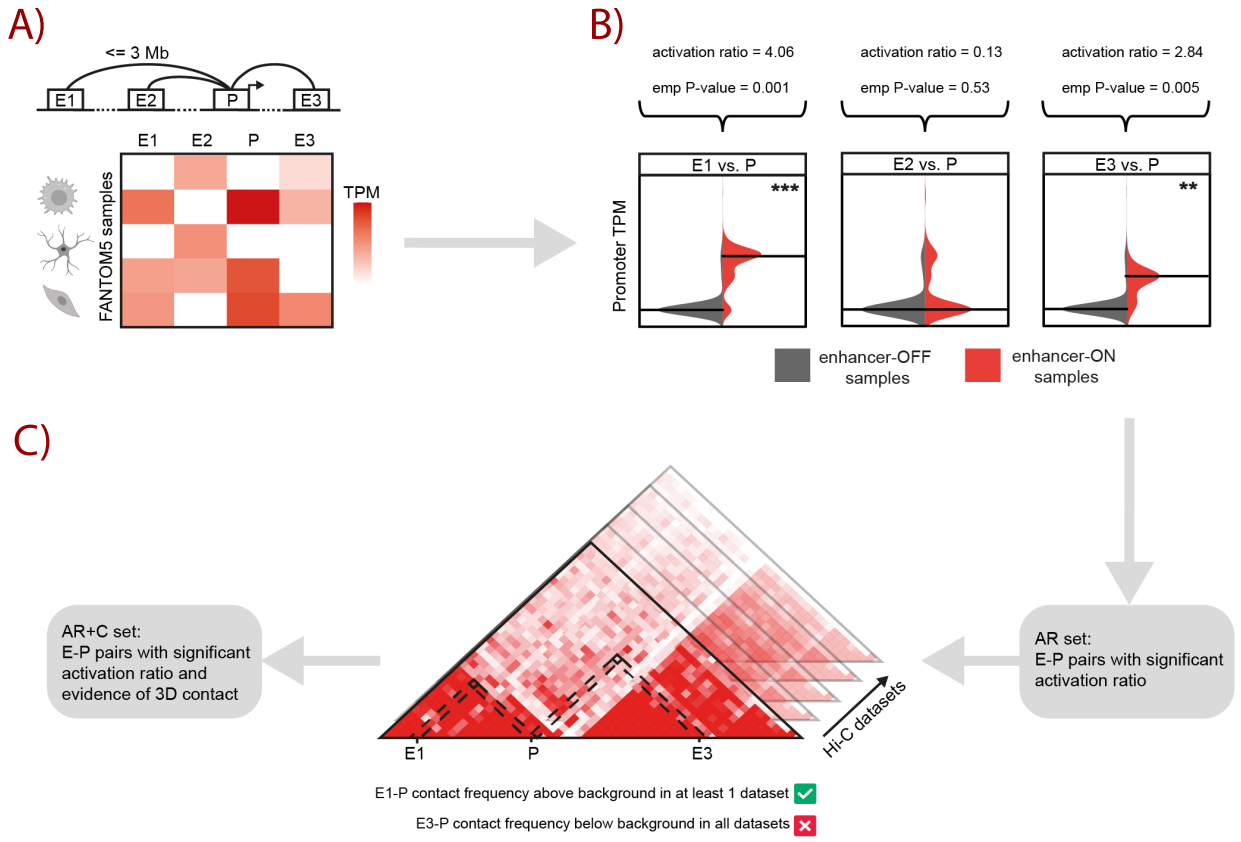


Figure 1.4: Overview of the AR+C.
A): an example showing the expression of 3 enhancers and 1 promoter (rows) across different tissues (columns) within a given 3Mb region. **B)**: three examples of how the expression of enhancers (E1-3) are co-expressed with the same promoter. In each sub-panel, the beanplot shows the distribution of promoter expression values (TPM) in tissues where the enhancer is off (grey), and one for the enhancer on (red). The effect size measure, or activation ratio, was calculated as the logarithm of the ratio of median promoter expression in enhancer-*on* samples over median promoter expression in enhancer-*off* samples. **C)**: significant E-P interactions from steps 1 and 2 are checked against Micro-C interaction data, and only E-Ps with contact frequency scores above a predefined threshold are kept. *Figure reproduced modified, with permission from Georgieva, 2022*

Ratio plus Contact, or AR+C. The AR+C was developed building onto the statistical framework in Barešić et al., 2020, and utilising CAGE data (see section 1.1.3.1 for a description of the CAGE method), generated by the FANTOM5 consortium (Andersson et al., 2014). The AR+C was built by evaluating the coordinated transcription of enhancers and promoters within 3 Mb windows, using FANTOM5 CAGE-defined enhancers and promoters in approximately 800 human samples. This method used a statistical measure called the *activation ratio* or *effect size* to assess the significance of each enhancer-promoter (E-P) pair. The effect size measure, or activation ratio, was calculated as the logarithm of the ratio of median promoter expression in enhancer-*on* samples over median promoter expression in enhancer-*off* samples.

The associations were further refined using enhancer-promoter 3D contact frequencies from high-resolution Hi-C datasets (see section 1.1.2.2 for a description of Micro-C) to generate a high-confidence set of E-P associations, referred to as *activation ratio+contact, or AR+C* in short. This method has been shown to identify biologically relevant regulatory associations, including long-range E-P interactions that are not detected, or are mis-assigned, by other methods (Georgieva, 2022). The method is summarised in Figure 1.4, and further details are offered in section 2.2.1.

1.3.1 Comparison of the AR+C with existing methods

Georgieva, 2022 benchmarked AR+C enhancer-promoter pairs against existing methods, such as the *ABC method* (Fulco et al., 2019; Nasser et al., 2021) and the closest gene association method, using promoter capture Hi-C data (Jung et al., 2019). While the AR+C performed equally or worse than existing methods for short distance enhancer-promoter associations, it showed improved accuracy when predicting long-range enhancer-promoter interactions in comparison to other standard approaches, such as closest gene assignment or the ABC method, as shown in Figure 1.5.

Further, the AR+C performs better than any existing methods at differentiating long-distance enhancer-promoter regulatory interactions in developmental genes, which

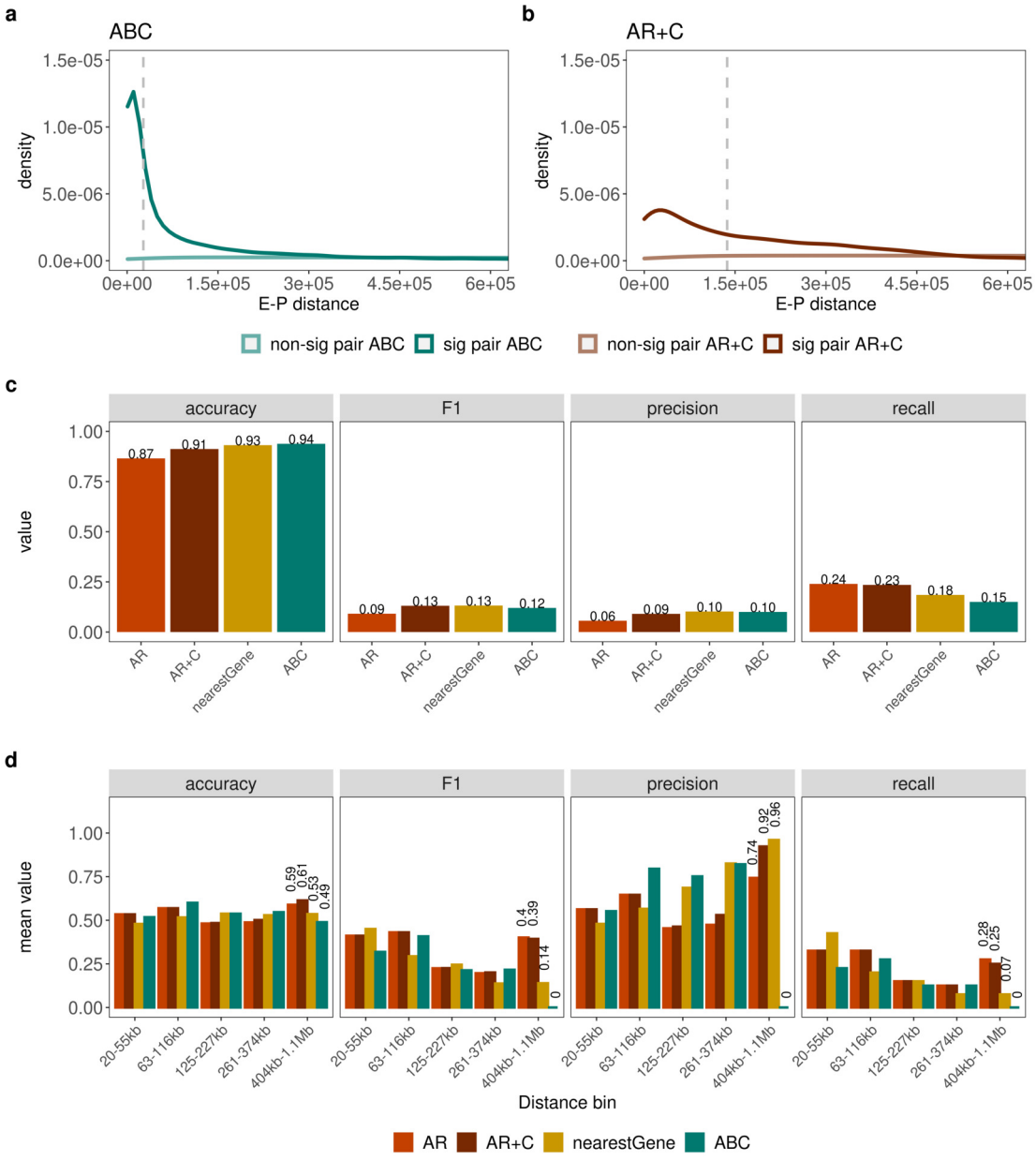


Figure 1.5: Comparison of the performance of the AR+C with other E-P assignment methods.
This used cardiomyocyte-specific PCHi-C loops as a benchmark (200 positive and 6904 negative loops).
The density plots show the linear distances for significant and non-significant cardiomyocyte-specific enhancer-promoter (E-P) pairs predicted by a) ABC, and b) AR+C. Grey dotted lines represent the median E-P distance for significant E-P pairs identified by each method.
(c) Accuracy, F1, precision and recall measures for cardiomyocyte-specific AR, AR+C and ABC predictions, as well as nearest-gene predictions.
(d) Same as in c), but PCHi-C loops were binned according to the linear distance spanned by each loop.
Figure reproduced from Georgieva, 2022, with permission.

are critically important for this work (see section 2.1.1). In fact, the AR+C showed high performance measures when used to classify known GRB targets, bystanders and non-GRB control genes, as compared to existing gold-standard classifications (Tan, 2017). Further, Georgieva, 2022 proceeded to develop a novel Random Forest-based machine learning model to uncover hundreds of previously uncharacterised GRB target genes (see section 2.2.1.1). Predicted target genes were shown to be enriched in specific gene ontology terms involving the synapse. The top biological process gene ontology (GO) terms significantly enriched included 'pattern specification processes', 'axon development', 'axonogenesis', 'embryonic organ development', 'forebrain development', while the top cellular component GO terms significantly enriched included 'transcription regulator complex', 'synaptic membrane', 'glutamatergic synapse', and several other components of the synapse. Importantly, predicted target genes were significantly associated with disease ontologies related to mental health disorders, neurodevelopmental disorders, such as Rett syndrome and intellectual disability (Georgieva, 2022).

In conclusion, the non-coding genome represents over 98% of the human genome, and it hosts a wealth of genomic regulatory elements (section 1.1.1). We have then reviewed the basic principles of 3D genome organisation, and how this helps to regulate gene expression through several levels of 3D folding (section 1.1.2). TADs delimit regions up to a few megabases long, within which most regulatory chromatin interactions seem to happen (section 1.1.2.1). Promoters and their interactions with enhancers are essential to start gene transcription, as well as to regulate it in a time- (ontogeny) and space- (cell-type) specific manner (section 1.1.3). Most enhancer-promoter interactions appear to happen within the same TAD, linking 3D structure and gene expression regulation. We have also seen that the genome is interspersed with conserved non-coding elements called CNEs, which are often sites of enhancer transcription (see section 1.2.1). CNEs tend to cluster, forming GRBs, which are structures where a target gene is tightly regulated by one or more enhancers, usually within the same TAD. GRBs are particularly important for the regulation of gene expression during development (see Figure 1.3 and section 1.2.2). Finally, I have described a resource

that has allowed to map enhancers to promoters genome-wide, by taking into consideration enhancer-promoter co-expression as well as the likelihood of 3D genome interactions. This resource is called AR+C and will be instrumental in annotating enhancers, which will form the core of this work. The AR+C is summarised in Figure 1.4, and in sections 1.3 and 2.2.1.

1.4 Complex disorders and Genome-wide association studies (GWAS)

1.4.1 Studying complex human disease genetics

Most of the several billion human DNA nucleotides can vary between individuals. If a single base varies in a population, it is called a single nucleotide polymorphism (SNP). Each SNP is annotated based on its position, and on the variation that it encodes. For example, SNP rs10000 is positioned on chromosome 7, and encodes base 5,973,522 of that chromosome on the 38.p13 genomic release. This SNP encodes an A > G variant (National Centre for Biotechnology Information, 2023). Each person has two homologous copies of each chromosome, and therefore can have three common combinations of rs10000: major/major alleles or 0 (AA in this example, also called phenotype AA), major/alternative alleles or 1 (AG in this example, also called phenotype Aa), or alternative-alternative alleles or 2 (GG in this example, also called phenotype aa); at the same position there might also be a rare A > T variant (Shastry, 2002). SNPs are not the only possible way DNA can vary between individuals: DNA can undergo deletions, duplications, as well as copy number variations, however these other types of variation will not be covered in this work as outside of our scope.

Each SNP can also be further characterised by other information, such as the minor allele frequency or MAF, which is the population frequency of the minor, or less common, allele (G in this case); or by the likely consequence of its mutation if it is a coding SNP, for example if it is *synonymous* (likely not to change the sequence of a protein) or *non-synonymous*

(likely to be affecting protein sequence, and possibly more damaging). Most SNPs, however, fall in non-coding regions and are therefore neither *synonymous* or *non-synonymous*, just *non-coding* SNPs.

1.4.1.1 Complex disorders

Some human traits and disorders are caused by a single, or by very few genetic mutations, which usually affect coding genes. Examples of these are cystic fibrosis or phenylketonuria. These are usually conditions that manifest early in life, and are at the severe end of the spectrum, as they produce abnormal proteins which directly cause the disease (Strachan & Read, 2018). The inheritance pattern for these conditions is *Mendelian* – one or more mutations to one specific gene can cause the disorder.

Most human physiological as well as pathological traits and conditions, however, are defined as *complex* in genetic terms. This category includes most psychiatric disorders, as well as many traits and disorders that are commonly studied, from body mass index, to blood pressure, to the risk of cardiovascular disease. Complex disorders, as the name implies, do not have an obvious genetic origin, and often the cause of the condition is not exclusively genetic. Therefore, the first step when assessing a disorder, is to study its heritability, or, in other terms, how much of the disease is genetically informed (see Polderman et al., 2015 and section 1.5.2). Heritability has been most often assessed through twin studies, which, by looking at the different prevalence of a condition between concordant and discordant twins, both fraternal and monozygotic, allow to apportion the amount of variance explained by genetics (Polderman et al., 2015). By subtraction, if the heritability for a condition is, for example, 50%, the condition will also show susceptibility to non-genetic, often called *environmental*, external factors, which will account for the remaining 50% of the risk.

Over the years it has become clear that for most complex disorders there are no 'causative' genetic variants, but that most common polymorphic SNPs show an association with the disease that can be expressed by an effect size measure. This association can be

a positive risk, a negative risk ('protective' variant), or neutral (non-associated variant) (Jostins & Barrett, 2011). Crucially, for most conditions there might not be any 'necessary' variants, with the risk (often called *susceptibility*) for a condition varying in a continuous fashion across the population.

1.4.1.2 Candidate genes and Linkage analyses

After studying a condition's inheritance pattern through twin- and pedigree-based studies, researchers started to investigate techniques to discover the individual molecular bases of human disorders and traits. The first approach to studying disease genetics was taken straight from Mendelian genetics: the so-called **candidate gene approach**. In this approach, one has to generate prior hypotheses about which genes might associate with a phenotype (e.g., if one knows of a gene that controls weight, they might hypothesise that the same gene might play a role in obesity). One then can test if SNPs in or around these candidate genes are associated with the condition (Tabor et al., 2002). This approach, however, is slow and expensive, and requires prior knowledge about the condition's genetics, something that often is not available.

A further approach to studying the association between disorders and genes was that of **linkage analyses**. Genetic linkage analysis consisted of studying the segregation of a trait of interest, such as a disorder, across members of several families – for each family, therefore, one needed multiple related family members available, both affected and unaffected. By genotyping genetic markers and studying their segregation through pedigrees, it was possible to infer their position relative to each other on the genome. This analysis allowed to find the genetic basis of many Mendelian disorders (Dawn Teare & Barrett, 2005).

However, most human traits and disorders are *complex*, and the initial approaches I have described are not well suited to studying complex disease genetics, due to their small scale, and the large number of putative variants involved in complex disorders. A large body of work has consequently gone into developing higher-throughput methods. These, together with a more thorough understanding of linkage disequilibrium, have en-

abled large-scale discovery. I will be introducing the next generation of studies, called genome-wide association studies, in the next paragraph, and the concept of linkage disequilibrium in section 1.4.3.

1.4.2 Genome-wide association studies (GWAS)

A genome-wide association study (GWAS) is an observational genetic study that aims to identify SNPs – or potentially other genetic variants not covered here – that are associated with a particular trait (which can be continuous, such as blood pressure) or disease (such as schizophrenia). For the purpose of this work I will focus on disease, or case-control, GWASes. Case-control GWASes typically involve comparing the genomes of individuals with a particular phenotype to those without the phenotype, looking for genetic variants that are more common in one group than the other. In a GWAS, researchers typically analyse hundreds of thousands or even millions of SNPs across the entire genome of individuals in both groups.

In GWAS various statistical techniques are employed to analyse the data (Balding, 2006). However, a common challenge is the uncertainty regarding the mode of inheritance (e.g., dominant, recessive, additive). Some methods are referred to as genetic model-free because they do not require pre-specification of a genetic model of inheritance. In contrast, other methods make such assumptions *a priori*.

1.4.2.1 GWAS penetrance functions and models of inheritance

Penetrance – in genetic terms – represents the proportion of carriers of a certain genotype (e.g., a specific allele of a SNP) showing a characteristic phenotype. Given the three potential configurations at a specific locus ($j = 0, 1, 2$), according to Gong et al., 2010 the probabilities of being affected by disease D depending on genotype g_j, which represent the disease penetrance for the same genotype, can be expressed as:

$$f_j = P(D|g_j) \tag{1.1}$$

For Mendelian disorders, $f_0 = 0$, or the risk of disease with no risk genotypes is zero. This of course is not the case for *complex* disorders, where the risk, as discussed, lies on a continuum, and therefore $f_0 > 0$. In the case of *complex* disorders, the risk of genotypes 1 and 2 can be expressed as the **relative** risk of disease D for an individual with g_1 or g_2 over that of an individual with g_0. In Bagos, 2013 these are defined as:

$$\gamma_1 = \frac{f_1}{f_0} \quad \text{and} \quad \gamma_2 = \frac{f_2}{f_0} \tag{1.2}$$

With these definitions in mind, one can then consider how to best model the risk of a genetic variant for a given condition. Table 1.1 shows an example of how to tabulate the distribution of cases and controls for an example disease D in a traditional GWAS at a single biallelic locus:

	g_0	g_1	g_2	total
cases	r_0	r_1	r_2	r
controls	s_0	s_1	s_2	s
total	n_0	n_1	n_2	n

Table 1.1: Distribution of cases and controls in a traditional GWAS at a single biallelc locus.

In the scenario of a case-control study, and in a genetic-model free framework, one can examine the association between the rows and columns of the 2×3 contingency table 1.1 using the traditional Pearson's χ^2_2 statistic which is distributed following a chi-square distribution with 2 degrees of freedom (Balding, 2006; Langefeld & Fingerlin, 2007). One can then compute the odds-ratios using the summary counts of the contingency table (Chen & Chatterjee, 2007).

Another model-free option is to calculate the two odds-ratios that correspond to the comparison of the genotypes carrying the risk allele (i.e., g_1 and g_2) against g_0. Such tests can be performed in a logistic regression framework using case/control status as the dependent variable; for a person carrying the g_j genotype (with $j = 1, 2$), we consider g_0 as the reference category and create two indicator variables taking values $x_j = 1$ for g_j and 0 otherwise (Balding, 2006):

$$logit(P(D|g_j)) = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{1.3}$$

In the same way as one can calculate SNP-disease associations in a model-free framework, it is possible to calculate the same measures following specific inheritance models (Bagos, 2013). Fitting a logistic model similar to equation 1.3, but using the coding of $x_i = (0, 0.5, 1)$ for the genotypes, equates to using an **additive model** of inheritance:

$$logit(P(D|g_i)) = \alpha + \beta_i x_i \tag{1.4}$$

However, coding of $x_i = (0, 1, 1)$, results in a test under the **dominant model**, whereas a coding of $x_i = (0, 0, 1)$ corresponds to testing the recessive model of inheritance (Zheng et al., 2003). For each SNP, then, a regression model is fitted, and a β coefficient (coefficient of model fit), as well as an odds ratio (OR, calculated as the exponential of the β coefficient) and a *p*-value of association are generated. These coefficients represent measures of association between each SNP and the phenotype of interest (Bagos, 2013). It is apparent that this method involves thousands, sometimes even million of statistical tests (one per SNP tested). This has to be taken into account when considering the resulting *p*-values. These can either be multiplied by the number of tests (Bonferroni correction), or can be considered significant below a conventional threshold (often 5×10^{-8}) (Jannot et al., 2015).

Despite most variants detected in GWASes being non-coding, the research focus has been mostly on non-synonymous genetic variants, such as SNPs that result in amino acid changes in proteins, or regulatory variants that affect gene expression or protein function (van de Bunt et al., 2015), as these are usually the ones showing higher effect sizes of association with the condition, and the easiest to link to a gene, and therefore to interpret mechanistically. In this work I want to focus instead on non-coding variants, and particularly on those falling inside enhancers. For this reason, I will explore non-canonical inheritance models for GWAS, such as the dominant and recessive models, which might be more

relevant to rarer variants with potentially larger effect sizes. In the next section I will introduce the concept of linkage disequilibrium, which is crucial to both GWAS, and to polygenic risk scoring (which is a common application of GWASes, as I will discuss in section 1.5).

1.4.3 Linkage disequilibrium

When talking about association studies, a key concept is that of linkage disequilibrium, or LD. Linkage disequilibrium refers to the degree of (non-random) association between alleles at different loci (positions on a chromosome), within a population. In other words, LD measures the tendency for certain alleles at different loci to occur together (within a population) more frequently than expected by chance. Stretches of DNA in high LD, which are therefore likely to be inherited together, form haplotypes, which typically descend from proximity on a single, ancestral chromosome (Reich et al., 2001). Several factors, usually relating to a population's history, can increase average LD across loci: a small original population's size (founder or bottleneck effects), genetic drift (or stochastic variation), and population admixture (e.g., the mixing of individuals from sub populations that have different allele frequencies); on the other hand, LD blocks can be broken down by recombination, which breaks down ancestral haplotypes. For this reason, LD decreases in proportion to the number of generations since the LD-generating event (Slatkin, 2008). LD is measured using measures derived from equation:

$$D_{AB} = p_{AB} - p_A \times p_B \tag{1.5}$$

which is the difference between the frequency of haplotypes carrying the pair of alleles A and B at two loci (p_{AB}) and the product of the frequencies of those alleles (p_A and p_B) (Slatkin, 2008). The D measure has some unfavourable mathematical properties, such as including negative numbers, and not having a fixed range. Therefore, D' was introduced, which represents the ratio of D to its maximum possible absolute value, given the allele frequencies

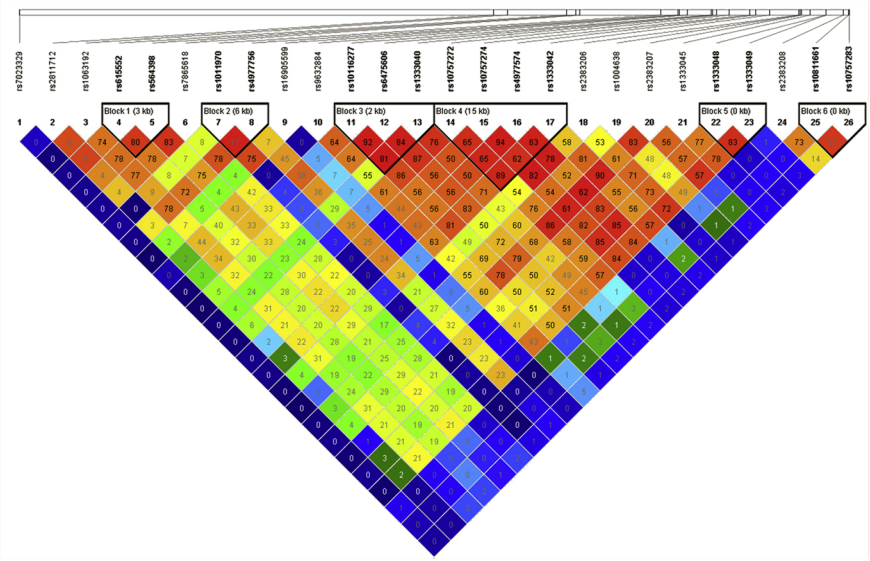


Figure 1.6: A block of LD on human chromosome 9p21.3.
A typical representation of LD between SNPs, as generated by the software Haploview.
The top represents a schematic section of a chromosome, with each *rs...* indicating a SNP.
In the heat-map at the bottom, each square represents the LD between two SNPs. Red squares indicate high LD (value close to 100, usually over 80). Blue squares indicate no linkage between SNPs.
Taken from Kalpana et al., 2019, copyright CC BY-NC-N.

(Lewontin, 1964). Another commonly used measure, derived from D, is r^2:

$$r^2 = \frac{D^2}{p_A \times (1 - p_A) \times p_B \times (1 - p_B)} \tag{1.6}$$

which ranges between 1 (full LD) and 0 (no LD) between loci (Slatkin, 2008). LD can be represented as in Figure 1.6, and is a key concept in disease-genetic marker association studies.

1.4.3.1 LD and genotypes: tagging SNPs and clumping

The majority of studies investigating human disease to date have been conducted using genotyping. Genotyping is a technique where researchers only type, or read, a finite number of SNPs from each study participant, instead of reading all SNPs (for example by whole genome sequencing). This is because, until very recently, sequencing a whole genome was extremely expensive, and therefore genotyping emerged as a technique to quickly and relatively inexpensively interrogate genetic variation genome-wide in a large number of

individuals (Syvänen, 2001).

Large-scale consortia such as the HapMap have mapped LD across human populations, producing public databases of human variation in relation to LD (McVean et al., 2005). This in-depth knowledge of LD has been exploited to make genotyping much more useful. In fact, randomly selecting a few thousand SNPs out of a whole genome would give very little information on the overall genetic picture of each study participant. However, the use of **tagging SNPs** has made the approach much more powerful. In a nutshell, a tagging SNP is a genetic variant that is used to represent a set of other genetic variants which are in high LD with it in a particular region of the genome. Therefore, tagging SNPs can capture the genetic variance of a whole region of interest. By genotyping a set of tagging SNPs, researchers can indirectly infer the genotypes of other variants in the same region that are not directly genotyped, a procedure called *genetic imputation* (Claussnitzer et al., 2020).

As we have seen, the human genome is comprised of haplotype blocks within which most SNPs are in high mutual LD, including polymorphic SNPs that are believed to be the basis of complex disease inheritance (see section 1.4.1.1). Another concept similar to tagging, and used to manage SNPs in LD with each other, is that of **clumping**. Clumping is performed after a GWAS, when the variants tested include SNPs that might be in LD with each other, which is the case with modern genotyping chips, as well as when working with imputed or sequenced data. Clumping involves grouping variants that are in LD with each other into 'clumps', and then selecting a representative variant from each clump for further analysis. A single representative variant is selected from each clump based on a pre-defined set of criteria – usually, the variant with the lowest GWAS p-value is selected (Marees et al., 2018). Clumping is used to calculate polygenic risk scores (see section 1.5 and Privé et al., 2019) in order to select LD-independent variants.

One of the downsides to using tagging SNPs and clumping to deal with LD, and to consequently reducing the number of loci considered, is that the results of a GWAS cannot be interpreted causally (Claussnitzer et al., 2020). A tagging SNP that is significantly associated with a condition of interest could in fact be the causal variant itself, however it could be

tagging a variable number of variants in LD with it. For this reason, in the past GWAS studies were often followed by so-called *fine mapping* studies, where the most significant regions of interest would either be sequenced in cases and controls, or else other *in silico* analyses could be performed to infer a putative causal variant. *In silico* analyses would exploit further information (annotations) on the region – which could entail for example regional DNA expression, classification, or other functional genetic information (Claussnitzer et al., 2020).

In this work, we want to use functional and tissue-specific annotations relating to enhancer-promoter interactions to further our knowledge of two complex disorders, schizophrenia and HCM.

1.5 Polygenic risk scores and heritability of human disease

1.5.1 Introduction to polygenic risk scores, or PRS

Polygenic-risk scores (PRSs) are an approach that allows to move from population-level analysis to infer an individual's liability to a specific condition. This is invaluable in terms of potential clinical applications. As described in section 1.4.2, GWASes measure the association of each SNP with a phenotype, producing a large list of effect size measures – typically, either a β coefficient or an odds ratio (OR) – for each SNP tested in the development sample. A polygenic risk score, or PRS, is a single value estimate of an individual's common genetic liability to a phenotype, calculated as the sum of their genome-wide genotypes, weighted by corresponding genotype effect size estimates derived from summary statistic GWAS data, as shown in equation 1.7 below (Dudbridge, 2013; Euesden et al., 2015). Usually, the largest available GWAS (the one with the biggest sample size) on the phenotype is used to calculate a PRS (Choi et al., 2020). There are more than one methods to calculate PRSs, however I will focus on the most widely used method, called the *classic PRS method*, or the *clumping and thresholding (C+T)* method, as this has been by a large margin the most widely used (Dudbridge, 2013).

PRS calculation involves a **base** GWAS, and a **target** population on which to cal-

culate the individual scores. There are a number of steps that are performed as part of PRS scoring, and which precede the adding up of the effect sizes. These steps are aimed at addressing intrinsic GWAS limitations owing to the fact that every GWAS is performed on a specific population (or more than one), and can therefore produce SNP effect size estimates that may not generalise well to other populations due to 'winner's curse' (where effects are overestimated in the initial discovery sample, leading to a subsequent decrease in effect size estimates in independent replication samples) and stochastic variation. LD among SNPs further complicates the aggregation of SNP effects across the genome (Choi et al., 2020). Therefore, key steps for calculating a PRS include:

- Adjusting (or shrinking) GWAS effect sizes for winner's curse;
- Tailoring the scores to the target populations;
- Dealing with LD.

I will be discussing each of these issues in the sections below.

1.5.1.1 Adjustment (shrinkage) of effect sizes

PRSs can generate poorly estimated results with high standard errors due to uncertainty in SNP effects and the fact that not all SNPs affect the trait being studied. To address this, two shrinkage strategies have been adopted: (A) shrinkage of effect estimates of all SNPs via statistical techniques such as LASSO, or Bayesian approaches such as in LDPred (Privé et al., 2020), and (B) the use of *p*-value selection thresholds as inclusion criteria for SNPs in the score, such as in the classic C+T method adopted in PRSice and used in this work (Euesden et al., 2015). The optimal shrinkage method is dependent on the mixture of null and true effect size distributions. The *p*-value selection threshold approach excludes SNPs with a GWAS association *p*-value above a certain threshold and includes only those below, effectively shrinking excluded SNPs to an effect size estimate of zero. Both methods involve tuning parameter optimization, and the optimal *p*-value threshold selected is within the context of forward selection ordered by GWAS *p*-value (Choi et al., 2020).

1.5.1.2 Controlling for linkage disequilibrium and dealing with target populations

GWAS association tests are typically performed per-SNP, which makes identifying the independent genetic effects very challenging due to strong correlations across the genome. To account for LD between SNPs, in the classical C+T method adopted in PRSice, the subset of SNPs with *p*-values lower than a specified GWAS threshold need to undergo clumping, a procedure which keeps only the top SNP per LD block (Choi & O'Reilly, 2019; Euesden et al., 2015). Clumping prioritises associated SNPs and retains multiple SNPs in the same genomic region if there are multiple independent effects there (for a more extensive discussion of clumping, see section 1.4.3). LD modelling requires estimation of LD between SNPs, and if the LD values derived from the base data are unavailable, then those from a closely matched reference sample, such as data from the 1000 Genome Project (Siva, 2008) are used to approximate these. However, if the base and target samples are drawn from different populations, then the PRS results may differ from those that would have been obtained had LD been computed in the base data itself.

1.5.1.3 PRS calculation

In summary, calculation of a PRS involves:

- Base and target QC, including excluding low quality imputed SNPs, rare variants (with MAF < 1%), removing individuals with high heterogeneity or missingness rates, and mismatching SNPs (those with different alternative alleles between the base and target datasets).
- Performing clumping, e.g. dealing with LD by selecting only the most significant variant from each LD block.
- **Calculating the PRS**. Using a target population of n individuals, to calculate the polygenic risk score (PRS) one has to add up the β for each SNP i out of a list of m SNPs, where $G_{i,j} = (0, 1, 2)$ is the genotype at SNP i for individual j. At threshold P_T, the

PRS for individual j can be calculated as:

$$PRS_{P_T,j} = \sum_{i=1}^{m} \beta_i G_{i,j} \tag{1.7}$$

- Performing thresholding, e.g. calculating the PRS at several *p*-value thresholds (e.g., all SNPs > 0.01, > 0.05, > 0.5, etc.
- Selecting the 'best-fit' threshold for the target population (if not using a pre-defined threshold – for example if comparing results with other PRSs at the same threshold).

For a detailed discussion of these methods, please see Methods section 3.3.

The PRS approach has been used to stratify people in the general population at increased risk of specific conditions, such as coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer (Khera et al., 2018). However, to date, no PRS has resulted in clinical usefulness in the field of psychiatry, despite the large GWAS sample sizes for conditions such as schizophrenia. This is due in large part to the very high polygenicity of this condition, as well as a degree of phenotype heterogeneity, which have resulted in a proportion of variance explained by the genetic factor for schizophrenia below 10% on the liability scale, out of an estimated heritability of around 80% (see Sullivan et al., 2003; Trubetskoy et al., 2022 and a further discussion of the issue in section 1.5.2). For this reason, in this work I have sought to increase the proportion of variance explained by the genetic factor for schizophrenia by taking into account non-coding genome annotations, and particularly long-range enhancer-promoter annotations, introduced in section 1.1.3.

1.5.2 Heritability and the missing heritability problem

There has been substantial debate around the proportion of genetic and environmental contribution to most human traits, from the simplest, such as eye colour, to the most complex phenotypes, among which most psychiatric conditions reside. A classic *twin design* study has been widely employed to unravel the extent to which genes and environment shape various human traits. By comparing the trait resemblance of monozygotic and dizy-

gotic twin pairs, we can gain insights into the relative contributions of these factors. While genetically speaking monozygotic twins are identical, dizygotic twins are full siblings (Polderman et al., 2015).

Many complex disorders (i.e., non-mono/oligogenic, and where multiple genetic variants interact with predisposing/precipitating environmental factors, see section 1.4.1.1) show high degrees of heritability (h^2), as calculated from twin studies (Polderman et al., 2015). However, the amount of genetic liability that can be attributed to the same conditions using common single nucleotide polymorphisms or SNPs from GWAS (h^2_{SNP}) is consistently much smaller; this has been called the 'missing heritability' problem (Maher, 2008). Figure 1.7 shows the degree of heritability (h^2), as compared to h^2_{SNP}, for a number of mental health conditions. One can easily notice that there is a big mismatch between the two measures.

Schizophrenia h^2 has been variably estimated to amount to 64% (Lichtenstein et al., 2009), 79% (Hilker et al., 2018), and up to 81% (Sullivan et al., 2003). However, schizophrenia's h^2_{SNP} – that is, the proportion of variance in liability attributable to all measured SNPs, as calculated using *SBayesS* (Zeng et al., 2021) – is of just 24%, and it goes down to just 7.3% if using the polygenic risk score of SNPs with GWAS $p<0.05$ across ancestries, in the most recent, very large GWAS from the Psychiatric Genomic Consortium (PGC) (Trubetskoy et al., 2022). Similarly, HCM shows very high heritability; this condition used to be considered exclusively Mendelian, even if it is now recognised as a complex disorder, with rare pathogenic variants in cardiac sarcomere genes identified in ∼35% of cases (see Tadros et al., 2023 and section 1.6.2). However, common SNP-based heritability (h^2_{SNP}) for HCM has been estimated as 18.1-28.8% in meta-analysis (Tadros et al., 2021), and considerably less when considering only SNPs below a certain *p*-value threshold (as per our estimates in this work, see chapter 3.4).

This work aims therefore to narrow the gap between h^2_{SNP} and h^2 for schizophrenia, by taking into account functional gene regulation, as discussed further in chapter 2.1.1. HCM, with a potentially very different genetic structure (as we will see in section 1.6.2), will act as a sensitivity analysis, as we hypothesise that there is a difference in the

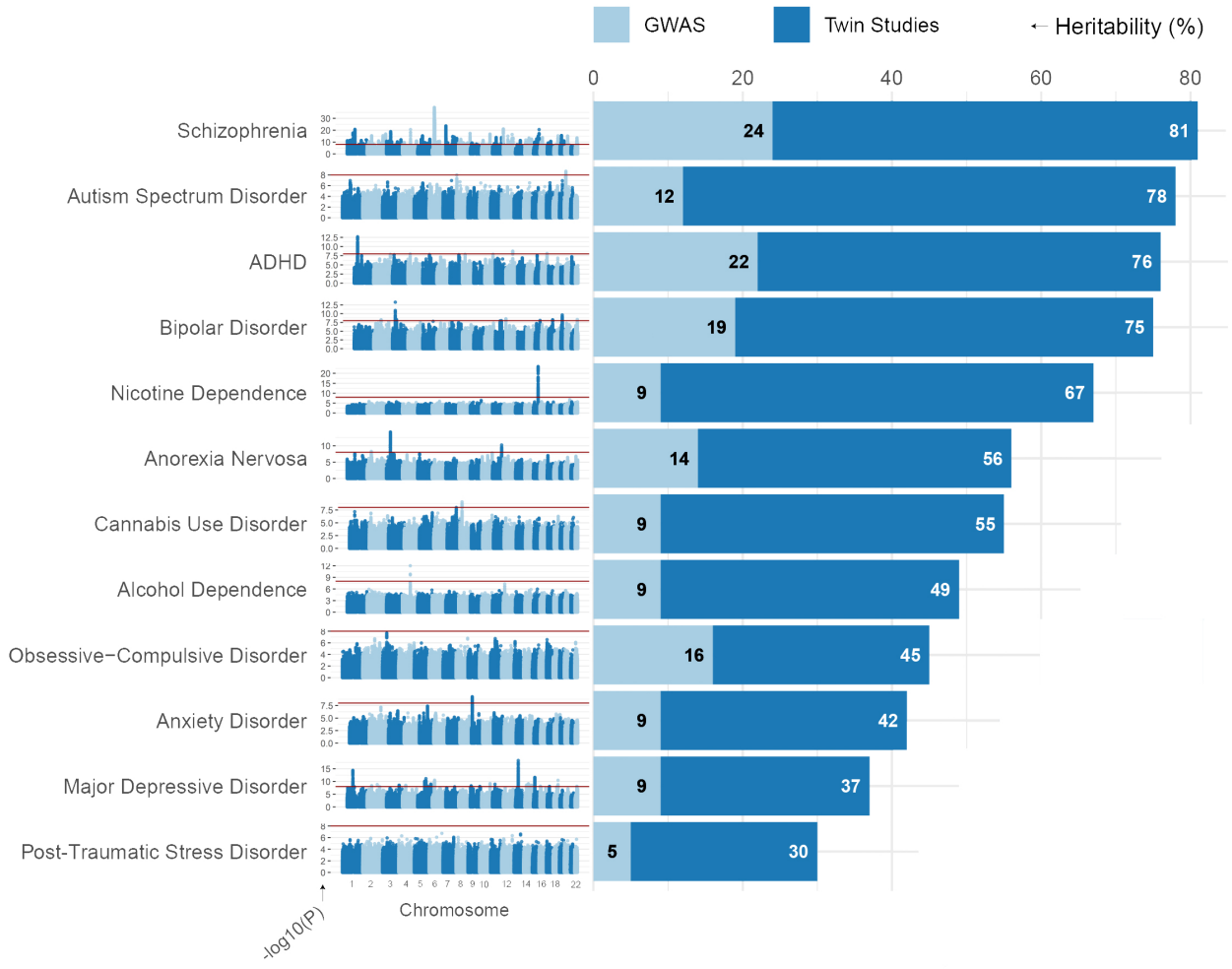


Figure 1.7: Heritability vs explained liability in select psychiatric conditions.
Genetic heritability for select mental disorders (dark blue), vs liability explained by common GWAS-derived variants (light blue).
Figure reproduced modified, with permission from Abdel Abdellaoui (who is also the copyright holder), from Abdellaoui and Verweij, 2022.

relevance of non-coding functional genome regulation between schizophrenia and HCM.

1.6 Introduction to model conditions: schizophrenia and hypertrophic cardiomyopathy

1.6.1 Introduction to schizophrenia

Schizophrenia is a severe mental illness affecting about 0.7-1% of the world population across their lifetimes (Saha et al., 2005). Schizophrenia has multiple clinical features, including behavioural, social, and biological manifestations, which fluctuate in one patient's lifetime, as summarised by McCutcheon et al., 2020 in Figure 1.8. The manifestations include positive symptoms, such as hallucinations – seeing, hearing or sensing things that are not really there – and delusions, or disordered thought – sometimes manifested for example as feeling that one is being persecuted without any objective evidence of that. Schizophrenia also has negative symptoms, which are less evident but can be even more damaging to a patient's life, which can include social isolation, emotional blunting, and reduced activity levels.

Schizophrenia is a significant burden to society, and incurs annual costs exceeding $150 billion in the United States (Cloutier et al., 2016), or £6.7 billion in England alone (Mangalore & Knapp, 2007). This is mainly due to the disorder typically manifesting in early adulthood, resulting in long-term impairments in social and occupational functioning. A diagnosis of schizophrenia also reduces life expectancy by up to 15 years – patients additionally have a lifetime risk of suicide-related death ranging from 5 to 10% (Hjorthøj et al., 2017).

Alongside the clinical picture, Figure 1.8 also describes some of the molecular features of schizophrenia. The condition is thought to be the result of the interaction of three factors: genetic liability, neurobiological factors, and external stressors. These have in turn been thought to be mediated by multiple inflammatory and cardio-metabolic changes – something that I have been investigating alongside others – in Osimo et al., 2018; Osimo et al., 2020a; Osimo et al., 2021a; Osimo et al., 2020b; Osimo et al., 2021b; Osimo et al., 2021c;

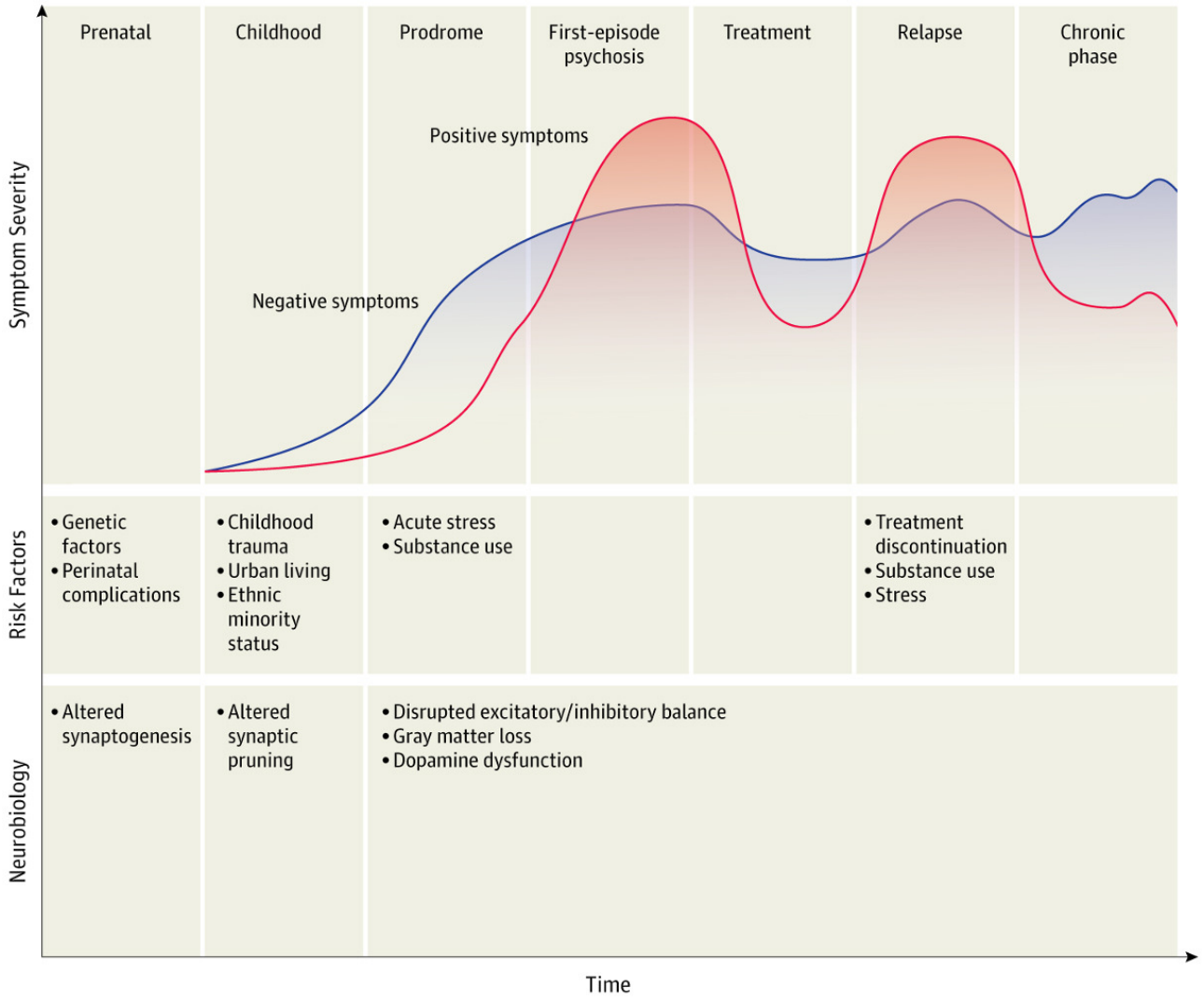


Figure 1.8: The Clinical Course of Schizophrenia.
People with schizophrenia usually show a fluctuating clinical picture, including relapses.
Figure reproduced with permission from: McCutcheon et al., 2020; Licensed from American Medical Association through RightsLink License Number 5507650799502.

Perry et al., 2021b; Perry et al., 2022; Pillinger et al., 2019b; Pillinger et al., 2019c.

In a genetically susceptible individual, the addition of one or more external risk factors, such as life adversity, infection, drug abuse as well as many others (reviewed in Radua et al., 2018) increases the risk of developmental abnormalities in the brain. Because of the spread of the timing of the risk factors for schizophrenia, which starts before birth and only ends when the condition manifests (Radua et al., 2018), as well as because of multiple findings from the field of neuropathology (Weinberger, 2017) as well as genetics (Trubetskoy et al., 2022), schizophrenia is often defined as a neurodevelopmental disorder – or a condition that manifests as the culmination of years of sub-clinical changes in the development of the brain (Osimo et al., 2019), immune system (Osimo et al., 2021b), and/or potentially other systems and organs (Pillinger et al., 2019a).

How to link genetic susceptibility and environmental factors to the complex manifestations of schizophrenia? At the level of neural transmission, most of the research has focussed on dopaminergic transmission, generating the **dopamine hypothesis of schizophrenia**. In brief, this states that multiple 'hits' interact to result in dopamine dysregulation, the final common pathway to schizophrenia. As shown in Figure 1.9, the endogenous and exogenous insults are thought to cause an increase in pre-synaptic dopamine function, which then cause the clinical phenotype of psychosis (Howes & Kapur, 2009). More recently, a 'synaptic model' of schizophrenia has been developed (Howes & Onwordi, 2023), suggesting that genetic and/or environmental risk factors render synapses vulnerable to excessive glia-mediated elimination triggered by stress during later neurodevelopment. The loss of synapses disrupts pyramidal neuron function in the cortex to contribute to negative and cognitive symptoms and disinhibits projections to mesostriatal regions to contribute to dopamine overactivity and psychosis. This model takes into account previous evidence on synaptic loss in schizophrenia (Onwordi et al., 2020; Osimo et al., 2019).

Consistent with the *dopamine hypothesis of schizophrenia*, most, if not all, current treatments for schizophrenia – called antipsychotic medications – rely on a single common biological pathway: dopamine receptor blockade in the brain (McCutcheon et al., 2019).

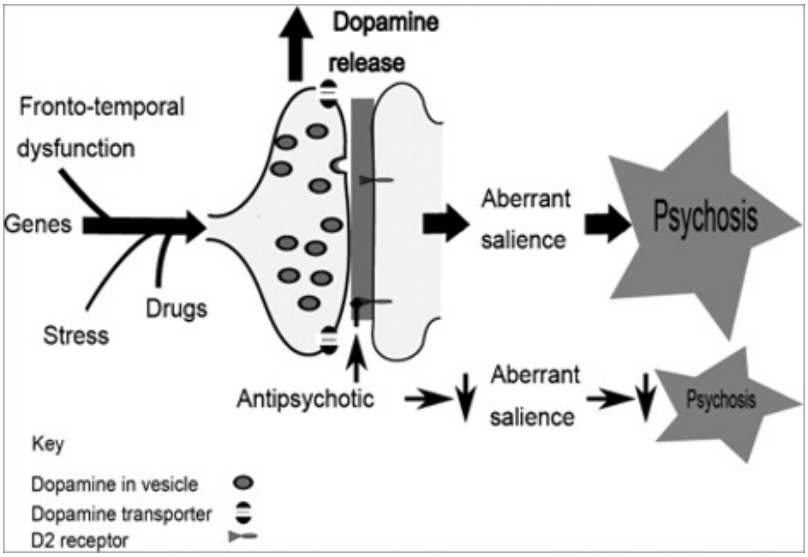


Figure 1.9: The Dopamine Hypothesis of Schizophrenia.
The final common pathway of schizophrenia. Predisposing and precipitating factors for schizophrenia converge on the presynaptic dopamine neuron, causing excessive dopamine release, which can cause psychosis, one of the core symptoms of schizophrenia.
Figure reproduced with permission from: Howes and Kapur, 2009. Licensed from Oxford University Press through RightsLink License Number 5510681501543.

These treatments are on average quite effective at treating positive and negative symptoms of schizophrenia (Lieberman et al., 2005), however they carry a high side-effect profile, which in turn contributes to high discontinuation rates. Newer, more tolerable treatments are urgently needed.

1.6.1.1 Schizophrenia genetics

In genetic terms, schizophrenia is highly heritable. Heritability estimates are of around 80% (see Sullivan et al., 2003 as well as section 1.5.2). However, as we have seen, only a small proportion of this heritability is explained by existing genetic studies of the condition. One of the reasons might be that schizophrenia's genetic burden is split between a large number of common variants with small individual relative risk, copy number variants, and rare coding variants (Coelewij & Curtis, 2018).

Schizophrenia genetics have been described as very highly polygenic, meaning that its inheritance relies primarily on a large number of common variants with small individual relative risk. As we have seen, genetic risk from common variants (usually defined

as SNPs with MAF $> 1\%$) can be studied through GWAS. This has required very large samples of cases and controls to power genetic association studies, which have been obtained through the formation of international consortia such as the Psychiatric Genetic Consortium. The latest in terms of PGC GWASes for schizophrenia was published in 2022, and it was a two-stage GWAS of up to 76,755 individuals with schizophrenia and 243,649 control individuals, reporting common variant associations at 287 distinct genomic loci of individually small effect (median odds ratio (OR) < 1.05) (Trubetskoy et al., 2022). Schizophrenia GWASes have not usually pointed to individual risk genes (with the exception of *C4*, which I will describe below), however the predicted genes affected by common variants have been studied through Gene Ontology. These pointed towards genes that are expressed in excitatory and inhibitory neurons of the central nervous system, but not in other tissues or cell types, as well as processes related to neuronal function, including synaptic organisation, differentiation and transmission, as is the case of the latest PGC GWAS (Trubetskoy et al., 2022).

Several rare (MAF $< 0.1\%$), recurrent copy number variants (CNVs) have also been robustly associated with schizophrenia, as exemplified by substantially higher rates of schizophrenia in carriers of 22q11.2 deletions (Karayiorgou et al., 1995; Marshall et al., 2017). A recent study has performed exome sequencing of a large number of patients and healthy controls (24,248 schizophrenia cases and 97,322 controls), suggesting that ultra-rare coding variants in a few dozens of genes might be conferring substantial risk for schizophrenia, with odds ratios between 3–50. Although these variants have large effects on risk in the individual, they make only a small contribution to overall heritability in the population owing to their rarity. The mutated genes had the greatest expression in central nervous system neurons and had diverse molecular functions, especially with regard to synaptic function (Singh et al., 2022).

Further, a small number of individual genes have been found to be implicated in schizophrenia: the first one to be discovered was *DISC1*. However, it has since proved difficult to elucidate the mechanism of this effect or to conclusively link other variants in

this gene to increased risk (Millar et al., 2000; Wang et al., 2018b). Several other individual genes were subsequently found to be associated with the condition, however for none there seemed to be any strong evidence for their necessity, or for their use to elucidate neural pathways to schizophrenia (Coelewij & Curtis, 2018).

A separate case is that of the *C4* locus. A significant schizophrenia GWAS hit was repeatedly found in the major histocompatibility locus (MHC), but often excluded from analysis because of the complex structure of MHC (Harrison, 2015). However, Sekar et al., 2016 set out to identify the specific gene or genes responsible for this association. This had proved difficult due to the complex and long-range linkage disequilibrium relationships observed in this region. Using droplet digital PCR, the group characterised complex variation in the C4 gene, which codes for complement component 4. They found that three types of variation commonly exist: the total number of copies can vary between 0 and 5; each copy can be long or short depending on whether it contains a HERV insertion; and each copy may be either of two paralogous genes denoted C4A and C4B. Four structural forms of C4A/C4B are commonly observed and each was shown to be associated with differing levels of expression. Using SNPs from the GWAS, it was possible to reliably impute these structural forms and show that predicted C4A expression from imputed C4 variants was associated with schizophrenia risk. The authors also demonstrated that C4A is present on neurons and synapses and postulated that increased expression could lead to increased synaptic pruning, which would produce the smaller number of synapses observed in patients with schizophrenia (Osimo et al., 2019). The study even demonstrated that mice lacking the C4 gene exhibited changes consistent with reduced synaptic pruning. While the effect on schizophrenia risk of C4 variants (in humans) is moderate, with an odds ratio of 1.3 between the highest and lowest risk structural forms, this paper is significant in that it links genetic variants to functional changes with a biologically meaningful impact (Sekar et al., 2016).

In conclusion, despite the large samples and huge efforts, involving GWAS of common variation, exome sequencing, candidate gene analyses, and studying at-risk populations such as subjects with 22q deletions, only a small portion of the disease variance is

explained by current genetic analyses (see section 1.5.2). In this work I aim to increase the proportion of schizophrenia heritability explained by the genetic factor by using enhancer annotations. I will set out specific hypotheses in section 1.7.

1.6.2 Introduction to HCM and its genetics

Hypertrophic cardiomyopathy is a complex cardiovascular disorder affecting about one in 500 of the general population (Tadros et al., 2021). Its relatively high mortality makes it the leading cause of sudden death in young people (Maron & Maron, 2013). As shown in Figure 1.10, HCM causes cardiac hypertrophy, an unhealthy increase in cardiac mass, which can cause a number of serious clinical consequences, including electrical abnormalities and consequent arrhythmias, which can lead to sudden death. The treatment of choice for severe cases is the implantation of a cardiac defibrillator, even if less severe cases can be treated pharmacologically, e.g. with β-blockers (Hamada et al., 2014).

In genetic terms, until a few years ago HCM was believed to be a Mendelian or oligogenic condition affecting the sarcomere – the contractile unit of the heart muscle cell (Maron & Maron, 2013) – but it has now been recognised as a complex disorder, with rare pathogenic variants in cardiac sarcomere genes identified in $\sim$35% of cases, and the remainder resulting from the interaction of several thousand SNPs genome-wide (Mazzarotto et al., 2020; Tadros et al., 2021; Tadros et al., 2023; Walsh et al., 2017).
HCM appeared a good candidate as a second model condition for this work because of the following reasons:

- The pathophysiology of HCM is likely to be highly tissue-specific, with most individual high-risk mutations found to date affecting the sarcomere, the contractile unit within the heart cell (Mazzarotto et al., 2020).
- The condition affects the heart, a tissue that is different and separate from the brain.
- As we will see in paragraph 2.1.1, developmental disorders appear to be affected by long-range gene regulation typical of genomic regulatory blocks, while conditions af-

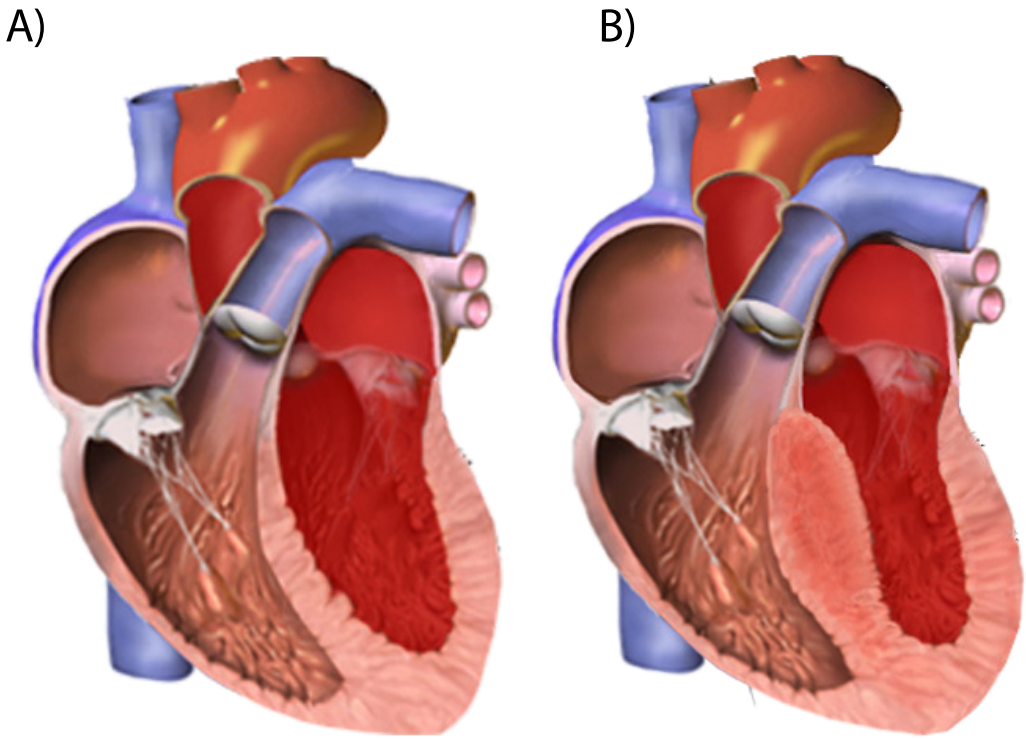


Figure 1.10: The normal vs hypertrophic heart.
A schematic representing a normal heart (A), vs the schematic of the heart of a patient with HCM (B). Note the absolute increase in left-ventricular wall thickness, particularly noticeable at the septum. *Modified from WikiMedia Commons, copyright CC BY-SA 4.0.*

fecting terminally differentiated tissues might be less likely to (Polychronopoulos et al., 2017). It is not known if the genetic susceptibility to HCM affects tissues across development, or if the polymorphisms are active mostly at the adult tissue level. Therefore, studying the effect of long-range regulation on HCM might act as a test of HCM's developmental credentials.

- Due to the availability of a yet unpublished genome-wide association study, the largest to date (Tadros et al., 2023) – and to the collaboration of members of the James Ware lab at the local LMS Institute.

1.7 Hypotheses and objectives

In this work I have set out to address the 'missing heritability' problem in schizophrenia, by exploring several avenues that could lead to increasing the variance explained by the genetic factor. I decided to use schizophrenia as my main model condition for several reasons, which I have discussed more extensively in section 1.6.1:

- Schizophrenia affects 1% of the world population, causing immense suffering, as well as a substantial burden and cost to society.
- There have been limited therapeutic breakthroughs, and the unmet need is great.
- Schizophrenia has a high genetic burden, which has been only partially explored to date. Genetic findings might both illuminate the pathophysiology of the condition, as well as lead to new therapeutic avenues.
- Schizophrenia is a neurodevelopmental condition. As we will see in paragraph 2.1.1, developmental disorders are expected to be affected by long-range gene regulation, and are therefore natural candidates for this work.
- As we have seen, the genetic susceptibility and environmental risk factors for schizophrenia are likely to converge on the neuron, potentially causing changes in struc-

ture (e.g., axon development), connectivity (e.g., pruning), or both, as the final common pathways. This allows to speculate that neural-tissue-specific genetic regulatory factors might play a role.

- Schizophrenia genetic risk from GWAS is particularly enriched in non-coding areas, some of which have no known functional significance. From the non-coding genome perspective, schizophrenia has been found to be enriched in promoter and enhancer variants associated with expression quantitative trait loci (eQTL) in the human brain (Roussos et al., 2014).

- GRBs have been shown to delimit the long-range regulatory landscapes of a number of developmental (or neuro-developmental) genes (Akalin et al., 2009). Schizophrenia genetic risk variants have been previously shown to preferentially fall within regions containing extremely conserved genetic elements and GRBs (see section 1.2.2 for a description of GRBs, and section 2.1.1 for an analysis of what they might mean for schizophrenia).

I have also selected a second model condition to compare to schizophrenia: hypertrophic cardiomyopathy, or HCM. HCM was selected as a comparison because of the very different genetic architecture (HCM has several high-risk variants, as well as many common ones, and it is thus less *polygenic*); because of its likely more pronounced tissue-specific aetiology (most high-risk variants affect the sarcomere, the contractile unit within the heart cell); and because of the lower likelihood of a developmental aetiology: in fact, while schizophrenia risk is known to originate from genetic and environmental factors spanning the pre-natal, to embryonal, to post-natal, to adult lifetimes, no developmental insults are known to increase HCM risk.

Using these exemplar conditions, in this work I aim to:

- **Chapter 2: Improve the understanding of the importance of enhancers and of GRBs, as well as study the relevance of tissue-specific enhancers for complex disorders**.

In this chapter I first develop neural and cardiac tissue-specific lists of enhancers significantly associated with at least one target gene, based on AR+C and FANTOM5 annotations. Then, I use partitioned LD score regression to test whether the common polygenic risks for schizophrenia and HCM are randomly distributed across tissue-specific enhancer- and GRB-based genomic partitions.

Here I also select the genomic partitions which are significantly enriched or depleted in GWAS signals, which I will take forward to the next analysis.

Hypotheses

– Neural-tissue-expressed enhancer genomic partitions and GRBs will be enriched in schizophrenia genetic heritability.

– Cardiac-tissue-expressed enhancer genomic partitions and GRBs will not be enriched in HCM genetic heritability.

- **Chapter 3: Develop *partitioned PRSs* as tools to increase explained disease heritability**.

Here I develop 'partitioned' polygenic risk scores, or PRSs where two genomic partitions (e.g., the *tissue-specific enhancers* and the *residual* partitions) are considered separately for polygenic risk scoring; enhancer-based SNPs are prioritised over genomic SNPs – and then disease heritability is calculated separately for the original GWAS for the condition (h^2_{SNP}), as well as for the partitioned PRSs (h^2_{pPRS}).

Here I also test if accounting for tissue-specific enhancer expression or target gene association measures – by multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by either the *effect size* of the tissue-specific enhancer, or by its tissue-specific expression – can increase disease h^2_{pPRS}.

Hypotheses

– h^2_{pPRS} will increase by fitting a model where a **prioritised** *tissue-specific enhancer*-based partition and a *residual* partition are separate predictors, as compared to the original GWASes for schizophrenia and HCM.

– *Tissue-specific enhancer* partition PRSs for schizophrenia and HCM will show higher coefficients of determination when accounting for tissue-specific enhancer expression or target gene association measures.

- **Chapter 4: Examine the importance of a dominant/recessive vs an additive genetic inheritance model in schizophrenia**.

  Most existing GWASes for complex disorders have utilised an *additive model* of inheritance – this assumes that the risk for each additional SNP is small, and that each additional allele acts independently by increasing risk. While most Mendelian diseases are classified as dominant or recessive, *nonadditive* effects have been seldom studied in the context of complex human disease heritability. On the other hand, *dominance effects* are central to the study of model disease fitness by population geneticists, and across organisms and conditions the average dominance of mutation of small effects should be approximately one-quarter (Manna et al., 2011). As a consequence, I have hypothesised that higher-priority variants (e.g., those falling inside tissue-specific enhancers for a relevant tissue) might follow a *nonadditive* (e.g., dominant/recessive) model of inheritance.

  In this chapter I will first perform an EP-WAS (association study of SNPs within enhancers) of schizophrenia using the dominant and recessive inheritance models in the UK Biobank, and then validate the findings in an external population (PGC cohort).

  Hypothesis

  Considering *nonadditive* (dominant or recessive) inheritance models for enhancer-based SNPs in schizophrenia will increase its h^2_{pPRS}.

Chapter 2

Schizophrenia and HCM heritability enrichment in tissue-specific enhancers

2.1 Introduction

2.1.1 Existing applications of the GRB model to psychiatric disorder genetics

In the general introduction chapters I have discussed the importance of the non-coding genome, particularly with regard to enhancer-promoter interactions for tissue- and time-specific gene expression regulation (section 1.1.3.2). We have also seen that GRBs can act as regulatory domains that delimit the span of long-range gene regulatory interactions (section 1.2.2). GRBs are characterised by a target-bystander gene structure, where one gene

is usually the target of multiple, conserved, non-coding medium- or long-range regulatory elements; other genes, termed bystanders, are usually unaffected (see Figure 1.3). GRBs are a useful conceptual model, as their presence at GWAS peaks can point to the gene that is most likely to be regulated, and thus, most likely to be affected, despite not being the closest gene to the lead GWAS SNP (Barešić et al., 2020).

As we have seen in section 1.6.1, schizophrenia is a neurodevelopmental condition characterised at the molecular level, among others, by alterations in synaptic pruning as well as by altered synaptogenesis (Howes & Onwordi, 2023; Onwordi et al., 2020; Osimo et al., 2019). Neurodevelopment, axon guidance, and synaptogenesis in particular, are some of the pathways that have been found to be under strong genetic regulation by enhancers following the GRB model (Georgieva, 2022), thus making studying schizophrenia in this context particularly relevant.

In fact, in previous research, Barešić et al., 2020 have found that GWAS-derived schizophrenia, autism spectrum disorder and bipolar disorder risk loci preferentially fall within GRBs. In addition to showing that neurodevelopmental conditions appeared to be enriched in loci under strong developmental regulation, this allowed to refine target gene predictions for non-coding disease-associated SNPs. In fact, these SNPs are often assigned to the closest gene, however Barešić et al., 2020 and Georgieva, 2022 have shown that there are much more likely targets within medium- and long-range enhancer regulation. The GRB model appears therefore useful as an additional tool when fine-mapping GWAS.

In this chapter I will be using LD score regression (LDSC) and cognate technique partitioned LDSC (see next section) to test if GRBs, as well as enhancer-based partitions, appear enriched in schizophrenia and HCM risk alleles.

2.1.2 Partitioned LD score regression (LDSC)

LD score regression is a technique that can be applied to GWASes, and which was developed to differentiate between polygenicity (or true effects) and confounding biases (Bulik-Sullivan et al., 2015). The polygenic model of complex disease inheritance postulates

that the cumulative burden of risk variants in a person determines their risk, which varies on a continuum (Visscher et al., 2021). Confounding biases, such as cryptic relatedness (i.e., kinship among the cases or controls that is not known to the investigator) and population stratification (genotype variations across populations unrelated to disease status), can yield an inflated distribution of test statistics in genome-wide association studies (GWAS) (Bulik-Sullivan et al., 2015).

Stratified, or partitioned, LD score regression is a technique cognate to LDSC, which allows to partition heritability across *functional categories* or genomic partitions (e.g., coding genes, promoters, enhancers, etc.), by only requiring GWAS summary statistics for a condition, and LD information from an external reference panel with ancestry matching the population studied in the GWAS (Finucane et al., 2015). This was initially developed to illuminate potential biological pathways underlying complex disorders. Finucane et al., 2015 found a large enrichment in conserved regions across many traits, and a very large immunological disease–specific enrichment of heritability in FANTOM5 enhancers.

2.1.3 Human enhancer annotations from the AR+C

As introduced in section 1.3, a method for the genome-wide identification of long-range regulatory associations was developed by Georgieva, 2022. This method will be further described at a technical level in section 2.2.1, however this involves measuring the coordinated transcription of FANTOM5 CAGE-defined enhancers and promoters, producing a measure of association strength (referred to as *activation ratio* or *effect size*), as well as a *p*-value for each enhancer-promoter pair. This is used to assess the coordinated transcription of CAGE-defined FANTOM5 enhancers and promoters within 3 Mb (independently of GRB boundaries) across human samples. Statistically significant associations between enhancers and promoters are further refined using 3D contact frequencies from five human high-resolution Hi-C datasets. The method is particularly useful for detecting long-range enhancer-promoter interactions, which have been shown to be important in genomic developmental regulation (Polychronopoulos et al., 2017).

The AR+C method utilises several human tissues from FANTOM5 for development, however – crucially – it does not address the question of which enhancers show a more tissue-specific expression pattern, and could therefore be more relevant for more tissue-specific conditions. For this reason, in this work I have initially created tissue-specific lists of enhancers from the initial global list of over $30K$ enhancers with at least one significant promoter interaction (significant co-expression and 3D contact). I also created separate 'control' lists, which either contained enhancers not expressed in the tissue of interest, or were not associated with any promoter. Subsequently, I used partitioned LDSC to test whether any of the tissue-specific enhancer-based lists were enriched in schizophrenia or HCM genetic heritability, using the largest GWASes to date for the conditions (Tadros et al., 2023; Trubetskoy et al., 2022).

2.2 Materials

2.2.1 Activation Ratio plus Contact (AR+C)

The AR+C used FANTOM5 CAGE data (see section 1.3 for a general introduction to the AR+C, and section 1.1.3.1 for a description of the CAGE technique and its relevance to enhancer discovery) to assess enhancers and promoters for coordinated transcription (independently of GRB boundaries) across approximately 800 human samples (Georgieva, 2022).

Enhancer-promoter associations were called by normalising the enhancer expression matrix for each sample from the FANTOM5 data hub. A promoter expression matrix was constructed by adding up the TPM values of all CAGE peaks overlapping a given promoter per sample. Samples were then selected based on quality criteria, excluding treated samples, and excluding those with unusually high or low expression levels. Enhancers and promoters with no expression across all samples were removed, resulting in a set of 241 cell line, 447 primary cell, and 120 tissue samples.

To identify significant enhancer-promoter pairs, a modified version of the method used in Barešić et al., 2020 was used. For each enhancer, all promoters within 3 Mb upstream

or downstream were considered as candidate targets, and the activation status of the enhancer was annotated as active or inactive in each sample based on its expression level. An activation ratio was defined as the log-fold change in median promoter expression in active versus inactive samples, and the statistical significance of the observed activation ratio was estimated based on a permutation test. Empirical *p*-values were then corrected for multiple testing using a 10% FDR correction, and only significant enhancer-promoter pairs with a positive activation ratio were considered. Finally, the highest expressed transcript for each protein-coding gene was used to simplify the analysis, resulting in 17,526 human promoters.

To refine the enhancer-promoter associations using chromosome conformation data, Hi-C and Micro-C datasets were downloaded from the 4D Nucleome Data Portal (Dekker et al., 2017) and were used to identify physical interactions between enhancers and promoters. The specific datasets used were:

- *Micro-C* data from HFF and hESC cells (Krietenstein et al., 2020), which is the highest-resolution human 3D contact dataset available to date.
- *In-situ Hi-C* data from GM12878 cells (Rao et al., 2014).
- *In-situ Hi-C* data from cardiac progenitor cells and ventricular cardiomyocytes (Zhang et al., 2019).
- *In-situ Hi-C* data from neurons and neural progenitor cells from PsychENCODE (Akbarian et al., 2015).

The data was downloaded in the `.hic` format and the `straw` tool was used to extract the contact matrix between chromosome regions of interest. The significant enhancer-promoter pairs identified earlier were then overlapped with the contact matrix, and those with significant contact were retained. The analysis resulted in a set of 3,447 human enhancer-promoter pairs with significant contact, which were further validated using a Hi-C validation method. The authors also performed a gene ontology analysis to identify enriched biological pathways associated with the validated enhancer-promoter pairs. See section 1.3.1 for a discussion of AR+C benchmarking against existing methods.

In summary, each E-P pair (where the enhancer and the promoter were within 3

megabases of each other) was associated with an effect size measure, called activation ratio, quantifying the median shift in promoter expression in tissues where the enhancer is active vs tissues where the enhancer is inactive, and its associated empirical *p*-value. Significantly associated E-P pairs were identified on a genome-wide basis by applying a 10% FDR threshold (see Figure 1.4 and Georgieva, 2022).

2.2.1.1 Human-mouse GRBs

A list of human-mouse GRBs based on CNEs with $> 98\%$ of homology was produced in Georgieva, 2022 using the same methodology as Harmston et al., 2017. Briefly, states of the human genome enriched for human-mouse CNEs ($> 98\%$ of homology over 50 bp, see section 1.2.1 for a description of CNEs) were identified using an unsupervised two-state hidden Markov model. These states were then merged together if they overlapped the same gene. CNEs that did not fall into an enriched state were discarded from the analysis. Adjacent CNEs were clustered into blocks. The resulting set of putative GRBs were those regions that contained at least 10 CNEs and a protein-coding gene, and were within a certain distance threshold from adjacent regions. The parameters used for the analysis varied depending on the evolutionary distance between the species being compared, and were empirically determined based on their ability to recapitulate known GRB boundaries (Georgieva, 2022).

2.2.2 PsychENCODE enhancers

As an external source of brain-specific enhancers – i.e., not derived from the AR+C – I also used a brain- (prefrontal cortex-) specific, high-confidence enhancer list from Wang et al., 2018a, obtained from http://resource.psychencode.org/; a resource called:

```
DER-03b_hg19 high_confidence_PEC_enhancers
```

Briefly, this is a list derived from the *Pyramidal resource* (Wang et al., 2018a), a collection of functional genomics data on the brain from multiple datasets, including PsychENCODE,

GTEx, ENCODE, and Roadmap Epigenomics. The data was uniformly processed and harmonized to create a dataset with a sample size of 1866 individuals. In addition to brain-cortex-specific enhancers, derived data products include brain-expressed genes, co-expression modules, single-cell expression profiles, and expression quantitative-trait loci.

2.2.3 GTEx Tissue specific expression quantitative trait loci (eQTLs)

Expression quantitative trait loci (eQTLs) are genomic loci that explain variation in expression levels of mRNAs. eQTL information was extracted from The Genotype-Tissue Expression (GTEx) project, which characterises genetic effects on gene expression levels across 44 human tissues (Battle et al., 2017).

2.3 Methods

2.3.1 Generation of tissue-specific enhancer lists

Enhancer lists were generated starting from FANTOM5 data (Andersson et al., 2014), as well as from the curated AR+C resource (Georgieva, 2022). Each enhancer's coordinates were increased by 100 bps (or less if close to another enhancer to avoid overlap) at each end to capture SNPs falling very close to, as well as within enhancers.

A first list includes all enhancers in E-P pairs with a positive effect size, evidence of significant E-P association, and 3D contact between Enhancer and Promoter – the list is called ALL SIGNIFICANT FANTOM5 ENHANCERS, with $N \approx 30K$. Further enhancer lists were generated, thus generating:

1. NEURAL/CARDIAC SIGNIFICANT ENHANCERS: a subset of ALL SIGNIFICANT FANTOM5 ENHANCERS, containing enhancers with positive tissue-specific expression (FANTOM5 neural or cardiac tissues average transcripts per million or tpm > 0).

2. NEURAL/CARDIAC SIGNIFICANT ENHANCERS WITHIN GRBS: a subset of NEURAL/CARDIAC SIGNIFICANT ENHANCERS, containing enhancers overlapping human-mouse GRBs over

98% of homology from either Georgieva, 2022 or Harmston et al., 2017.

3. BRAIN/HEART ENH-PROMOTER-EQTLS: a subset of ALL SIGNIFICANT FANTOM5 ENHANCERS, overlapping significant GTEx eQTLs for either brain or heart (see section 2.2.3). An eQTL was defined as significant in brain if the association between SNP and corresponding gene showed p-value < 0.05 in all of Brain Cortex, Brain Anterior cingulate cortex (BA24), Brain Frontal Cortex (BA9), and Brain Hippocampus; for heart, an eQTL was defined as associated with its corresponding gene if p-value $< 10^{-3}$ in both Heart Atrial Appendage and Heart Left Ventricle. 'Overlap' for this work is defined as an eQTL (a SNP-gene pair) where the SNP is contained within a significant AR+C enhancer, and the target gene is the same as the E-P pair.

4. NON NEURAL/CARDIAC ENHANCERS: Control list 1. A list of enhancers lacking tissue-specific expression (neural or cardiac) from FANTOM5 data – not necessarily significantly associated to a promoter.

5. NON ASSOCIATED ENHANCERS: Control list 2. Genome-wide enhancers with no significant target gene (either no co-expression, or 3D association with a promoter, or both).

The tissue-specific enhancer lists are individually described in section 2.4.1.

2.3.1.1 Tissue specificity of enhancer lists

Neural tissue specific enhancers were selected if they showed expression in any of the following FANTOM5 tissues/cells:

- Adult brain tissue from donor (1×);
- Adult brain tissue pool (1×);
- Neuronal primary cells from donor (3×);
- Occipital cortex tissue from donor (2×).

Induced pluripotent stem cell-derived neurons were excluded as showing noise in expression signals.

Myocardial tissue specific enhancers were selected if they showed expression in any of the following FANTOM5 tissues/cells:

- Cardiac myocyte primary cells from donor (3×)
- Adult heart tissue pool (1×)
- Fetal heart tissue pool (1×)

Fibroblasts, embryonic stem cells, mesenchymal cells, and valve tissues were excluded as incompatible with the likely aetiology of HCM.

2.3.2 Measurement of the relative importance of enhancer-based partitions for specific GWAS using partitioned LDSC

Stratified or partitioned LD score regression (LDSC) was introduced in section 2.1.2; it is a method introduced by Finucane et al., 2015 for measuring heritability enrichment for a specific condition (from GWAS summary statistics) for specific genomic partitions. In this work I used the LDSC software (Bulik-Sullivan et al., 2015) to test whether tissue-specific enhancer genomic partitions were enriched or depleted in significant disease-specific SNPs. This took several steps:

1. Reformatting GWAS summary statistics – for schizophrenia from Trubetskoy et al., 2022; and for HCM from Tadros et al., 2023; including re-stranding each SNP to match 1000 Genome Project (Clarke et al., 2012) reference alleles. This was done using the $munge\_sumstats.py$ script from Bulik-Sullivan et al., 2015.

2. Using the enhancer lists generated in section 2.3.1 above; annotating and re-stranding each SNP to match 1000 Genome Project (Clarke et al., 2012) reference alleles. This step made use of the $make\_annot.py$ script from Bulik-Sullivan et al., 2015.

3. Computing LD scores for each partition for each chromosome. This step was conducted using the $ldsc.py$ script from Bulik-Sullivan et al., 2015, with flags $--l2$ $--thin-annot$ $--ld-wind-cm$ 1.

4. Calculating partitioned heritability using the $ldsc.py$ script from Bulik-Sullivan et al., 2015, with flag $--h2$.

2.3.3 Software and code availability

Statistical analyses were performed in R 4.2.2 (R Core Team, 2023), using Tidyverse libraries (Wickham et al., 2019). Stratified LD score regression (LDSC) was performed using the LDSC package (Bulik-Sullivan et al., 2015; Finucane et al., 2015). The FANTOM5 dataset is publicly available through the *CAGEr* package (Haberle et al., 2015). AR+C annotations (including genome-wide enhancer-promoter associations) will be available once published by Georgieva et al, and may be made available upon reasonable request.

The code for this chapter is available on the GitHub repository at: https://github.com/emosyne/partitioned_LDSC.

2.4 Results

2.4.1 Tissue-specific enhancer partitions

Tissue-specific enhancer lists were generated as described in Materials and Methods, sections 2.2 and 2.3. In brief, I used enhancer annotations from the AR+C (Georgieva, 2022) and from FANTOM5 (Andersson et al., 2014), measuring enhancer-promoter associations in human and mouse genomes using Cap Analysis of Gene Expression sequencing (CAGE, see section 1.1.3.1) and chromosome conformation data from Micro-C (see section 1.1.2.2) experiments. The method involved identifying significant human enhancer-promoter pairs, and incorporating chromosome conformation data to refine the results (see section 2.2.1). Significant enhancer-promoter pairs had a measure of *effect size* available, quantifying the strength of the enhancer-promoter association, as well as an association p-value.

As described in methods section 2.3.1, the initial significant enhancers list (ALL

SIGNIFICANT FANTOM5 ENHANCERS) was filtered based on tissue-specific expression. A first group included enhancers with tissue-specific expression in neural or cardiac tissues from FANTOM5 data. Enhancers meeting this criterion were defined NEURAL/CARDIAC SIGNIFICANT ENHANCERS. A subset of these enhancers that overlapped a human-mouse GRB were called NEURAL/CARDIAC SIGNIFICANT ENHANCERS WITHIN GRBS. Significant enhancers were also overlapped with brain/heart GTEx eQTLs (see section 2.2.3), resulting in enh-promoter-eQTLs specific to the brain and heart tissues, which were called BRAIN/HEART ENH-PROMOTER-EQTLS or E-P_eQTLs. Control lists were generated, including enhancers that lacked tissue-specific expression (neural or cardiac) from FANTOM5 data and were not necessarily significantly associated with a promoter, called NON NEURAL/NON CARDIAC ENHANCERS. Additionally, a list included genome-wide enhancers with no significant target gene, called NON ASSOCIATED ENHANCERS. For the brain/schizophrenia component of the study only, these lists were also compared to a list of prefrontal-cortex specific enhancers with no information about promoter specificity from PsychENCODE (see section 2.2.2).

2.4.1.1 Neural tissue and schizophrenia

Table 2.1 describes the size, enhancer-promoter distance, and tissue specificity of the lists generated for, and used in this work, with regards to the neural-specific part. The first line is the list of all 30,049 enhancers with significant co-expression with at least one promoter, and with evidence of 3D contact between the enhancer and the promoter, from the AR+C (ALL SIGNIFICANT FANTOM5 ENHANCERS, from Georgieva, 2022). The enhancer and promoterd in this set were on average 182,806 bps apart (SD = 209,193). Of these, 21,145 enhancers showed significant neural expression, with a similar average enhancer-promoter distance of 189,985 bps. Of the neural-specific enhancers, 7,582 (35.9%) overlapped a GRB, and showed a slightly larger average enhancer-promoter distance of 223,044 bps (SD = 253,560). The 687 enhancers overlapping significant GTEx eQTLs for brain were set at a shorter average enhancer-promoter distance of 137,897 bps (SD = 175,495), as expected for

eQTLs.

Finally, the two negative association lists showed much larger distances between promoters and enhancers (both $>$ 800,000 bps), however this has little meaning as the promoters were not always significantly associated with the enhancer.

2.4.1.2 Cardiac tissue and HCM

Table 2.2 describes the size, enhancer-promoter distance, and tissue specificity of the lists generated for, and used in this work, with regards to the cardiac-specific sensitivity analysis. Once more, the first line is the list of all 30,049 enhancers with significant co-expression with at least one promoter, and with evidence of 3D contact between the enhancer and the promoter from the AR+C (ALL SIGNIFICANT FANTOM5 ENHANCERS, Georgieva, 2022). Of these, 3,126 overlapped a GRB, and showed a slightly larger average enhancer-promoter distance of 202,973 bps (SD = 230,571). The 905 enhancers overlapping significant GTEx eQTLs for heart were set at a shorter average enhancer-promoter distance of 125,747 bps (SD = 155,268). Finally, the two negative association lists showed much larger distances between promoters and enhancers (both $>$ 900,000 bps), however this has little meaning as the promoters were not always significantly associated with the enhancer.

2.4.2 Tissue specific enhancers and heritability for schizophrenia and HCM

The aim of this section is to measure the amount of heritability for schizophrenia and for HCM coming from tissue-specific enhancer lists, and to test whether this exceeds the heritability that would be expected by chance. To do so, I utilised LDSC; see Methods section 2.3.2 for details. In brief, each of the tissue-specific enhancer lists was treated as a genomic partition, and compared with existing genomic partitions as previously done in Finucane et al., 2015. GWAS summary statistics – namely the schizophrenia GWAS by Trubetskoy et al., 2022 and the HCM GWAS by Tadros et al., 2023 – as well as information about genomic partitions, were studied through the LDSC software. This provides an estimate of the heritability enrichment for each genomic partition, as compared to all GWAS SNPs. In LDSC

Table 2.1: Neural/brain-specific enhancer lists: size, enhancer-promoter distance, and tissue specificity

Enhancer list	Number of enhancers	Mean E-P distance	10^{th}, 50^{th}, 90^{th} centile of E-P distance	Mean (SD) promoter tissue specificity index	10^{th}, 50^{th}, 90^{th} centile of promoter tissue specificity index
ALL SIGNIFICANT FANTOM5 ENHANCERS	30,049	182,806	10,271, 113,450, 440,261	0.51 (0.25)	0.21, 0.45, 0.9
NEURAL SIGNIFICANT ENHANCERS	21,145	189,985	11,171, 121,261, 450,995	0.51 (0.25)	0.21, 0.45, 0.91
NEURAL SIGNIFICANT – WITHIN GRBS	7,582	223,044	13,008, 146,946, 513,367	0.54, (0.26)	0.23, 0.49, 0.93
NEURAL SIGNIFICANT – NOT IN A GRB	13,563	170,971	10,425, 108,561, 417,490	0.5 (0.25)	0.21, 0.44, 0.89
BRAIN ENH-PROMOTER-EQTLS	687	137,897	7004, 68,795, 375,360	0.49 (0.23)	0.22, 0.44, 0.86
NON NEURAL ENHANCERS	20,357	824,099	21,802, 380,107, 2,382,626	0.54 (0.23)	0.24, 0.53, 0.89
NON ASSOCIATED ENHANCERS	34,560	904,139	27,611, 472,960, 2,440,948	0.5 (0.23)	0.22, 0.45, 0.86

Table 2.2: Cardiac-specific enhancer lists: size, enhancer-promoter distance, and tissue specificity

Enhancer list	Number of enhancers	Mean E-P distance	10^{th}, 50^{th}, 90^{th} centile of E-P distance	Mean (SD) promoter tissue specificity index	10^{th}, 50^{th}, 90^{th} centile of promoter tissue specificity index
ALL SIGNIFICANT FANTOM5 ENHANCERS	30,049	182,806	10,271, 113,450, 440,261	0.51 (0.25)	0.21, 0.45, 0.9
CARDIAC SIGNIFICANT ENHANCERS	8,959	180,963	8921, 114,072, 439,971	0.45 (0.23)	0.21, 0.4, 0.83
CARDIAC SIGNIFICANT – WITHIN GRBS	3,126	202,973	8,920, 128,585, 475,988	0.46 (0.22)	0.21, 0.4, 0.82
CARDIAC SIGNIFICANT – NOT IN A GRB	5,833	168,439	9,006, 104,632, 417,681	0.45 (0.23)	0.2, 0.4, 0.83
HEART ENH-PROMOTER-EQTLS	905	125,747	6,644, 64,652, 342,553	0.48 (0.22)	0.21, 0.45, 0.81
NON CARDIAC ENHANCERS	40,870	997,494	33,989, 626,900, 2,520,033	0.51 (0.24)	0.23, 0.47, 0.89
NON ASSOCIATED ENHANCERS	34,560	904,139	27,611, 472,960, 2,440,948	0.5 (0.23)	0.22, 0.45, 0.86

the enrichment and associated p-value test the hypothesis that a genomic partition – e.g., a specific list of enhancers – shows a significant enrichment in heritability for a condition – e.g., schizophrenia – as compared to the genome's average.

2.4.2.1 Neural-tissue-specific enhancers and schizophrenia heritability

In this section, I tested several genomic partitions for enrichment in schizophrenia-associated SNPs; some of these regions were standard genomic annotations, such as coding genes (i.e., exons), introns or untranslated gene-flanking regions (3′ or 5′ UTRs). I then tested our regions of interest, i.e., enhancer-based partitions.

Standard genomic partitions

I first tested whether the main standard genomic partitions, as annotated from genomic data in Andersson et al., 2014; Hnisz et al., 2013; Hoffman et al., 2013, were significantly enriched in schizophrenia heritability, comparing them to a 'base' partition containing all GWAS SNPs. Figure 2.1 shows that genes, both in terms of coding exons, and in terms of non-coding regions (introns), were enriched in schizophrenia heritability, both showing Benjamini-Hochberg-adjusted p-values ≤ 0.01. Introns, in particular, hosted 36.9% of the SNPs (the largest partition of all), and 50.2% of the heritability. In terms of peri-genic regions, promoters and promoter-flanking regions showed positive enrichment values, which however did not pass the false discovery rate threshold of 0.05. 3′ untranslated regions were significantly enriched in schizophrenia heritability, with 1% of SNPs and 5% of heritability (adjusted p-value = 0.03); schizophrenia heritability enrichment of 5′ untranslated regions was not significantly different from the null.

Given the focus of this work on enhancers, I included four different classifications for enhancers (Andersson et al., 2014; Hoffman et al., 2013), weak enhancers (Hoffman et al., 2013), and super-enhancers (Hnisz et al., 2013). Super enhancers were the largest partition, including 16% of SNPs, and were significantly enriched in schizophrenia heritability, with 24.7% (BH-adjusted p-value < 0.001), while all other categories of enhancers were not

significantly enriched or depleted in schizophrenia heritability (Figure 2.1).

Tissue-specific and control enhancer genomic partitions

In this section I tested whether enhancer-based partitions, as well as a human-mouse GRB partition, were enriched in schizophrenia heritability. As described in section 2.2.1.1, the human-mouse GRBs list with homology $> 98\%$ was produced in Georgieva, 2022 using the same methodology as Harmston et al., 2017. In this work we confirm that GRBs are significantly enriched in schizophrenia heritability, with a proportion of 31% of the tested SNPs falling by chance in a human-mouse GRB, compared to a partition heritability of 41.7% (adjusted p-value$<$0.001). See Figure 2.2.

To test whether enhancers are generally important for schizophrenia genetics, next, I tested enhancer-based partitions for schizophrenia heritability enrichment. Figure 2.2 describes the results; the first test refers to all FANTOM5 enhancers, significant or not (ENHANCER ANDERSSON; BH-adjusted p=1), which, as shown in the previous paragraph, were not enriched for schizophrenia heritability; neither were all enhancers with significant AR+C promoter interactions – unselected for tissue expression (ALL SIGNIFICANT FANTOM5 ENHANCERS; BH-adjusted p=1). As expected, FANTOM5 non-associated enhancers (e.g., those not associated to any nearby promoter) were not enriched for schizophrenia heritability either.

Testing tissue-specific lists, I found that, importantly, FANTOM5 promoters with no neural expression (NON NEURAL ENHANCERS in Figure 2.2; either significantly associated to a gene or not) were depleted in schizophrenia GWAS heritability, with an enrichment value of -8.4 (BH-adjusted p=0.045), and the only partition to be so. Of the neural-expressed partitions, the largest, including all $\sim$21k enhancers with significant AR+C promoter interactions (NEURAL SIGNIFICANT ENHANCERS) was enriched in schizophrenia heritability, with a value of 8.2, BH-adjusted p=0.01. A subset of this list, including $\sim$8k significant neural enhancers overlapping a GRB, was even more enriched, with an enrichment value of 29.9, and BH-adjusted p<0.0001. This contrasted with the $\sim$14k neural-expressed enhancers with

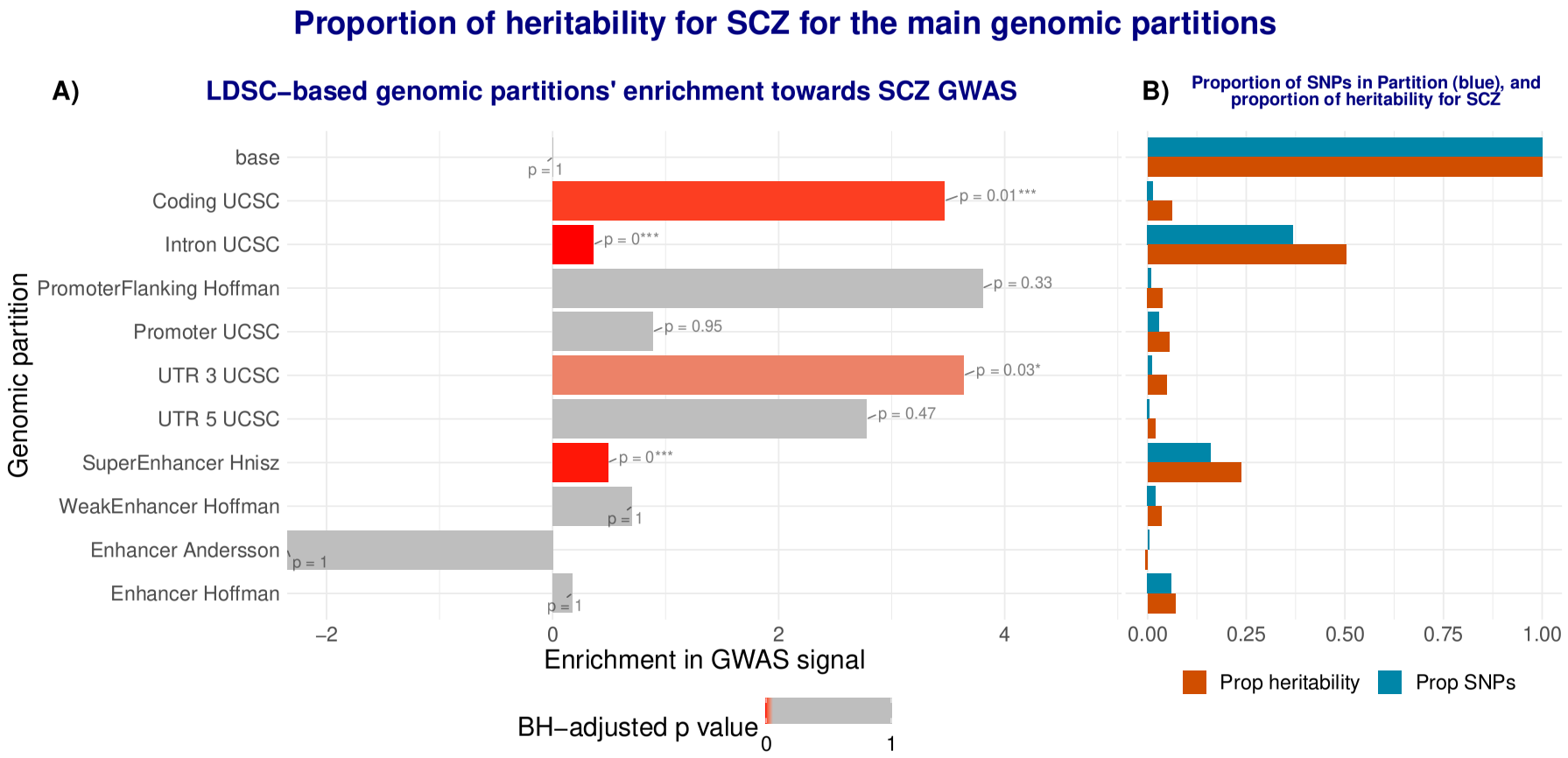


Figure 2.1: Partitioned LDSC for schizophrenia; main genomic partitions.
Partitioned LD score regression (LDSC) of the main genomic partitions for the latest schizophrenia GWAS (Trubetskoy et al., 2022).
Panel A shows the enrichment value – a positive value indicates more SNPs for schizophrenia in this partition than expected by chance, and a negative value indicates depletion. *P*-values are corrected for multiple testing using the Benjamini-Hochberg (BH) method.
Panel B shows the proportion of SNPs (Prop SNPs in blue) and of schizophrenia heritability for the SNPs in the partition (Prop heritability, in red) for each genomic partition.
Abbreviations: SCZ: schizophrenia; base: all GWAS SNPs; UCSC: University of California Santa Cruz database; UTR 3: 3' untranslated gene regions; UTR 5: 5' untranslated gene regions; Hoffman: (Hoffman et al., 2013); Andersson: (Andersson et al., 2014); Hnisz: (Hnisz et al., 2013).

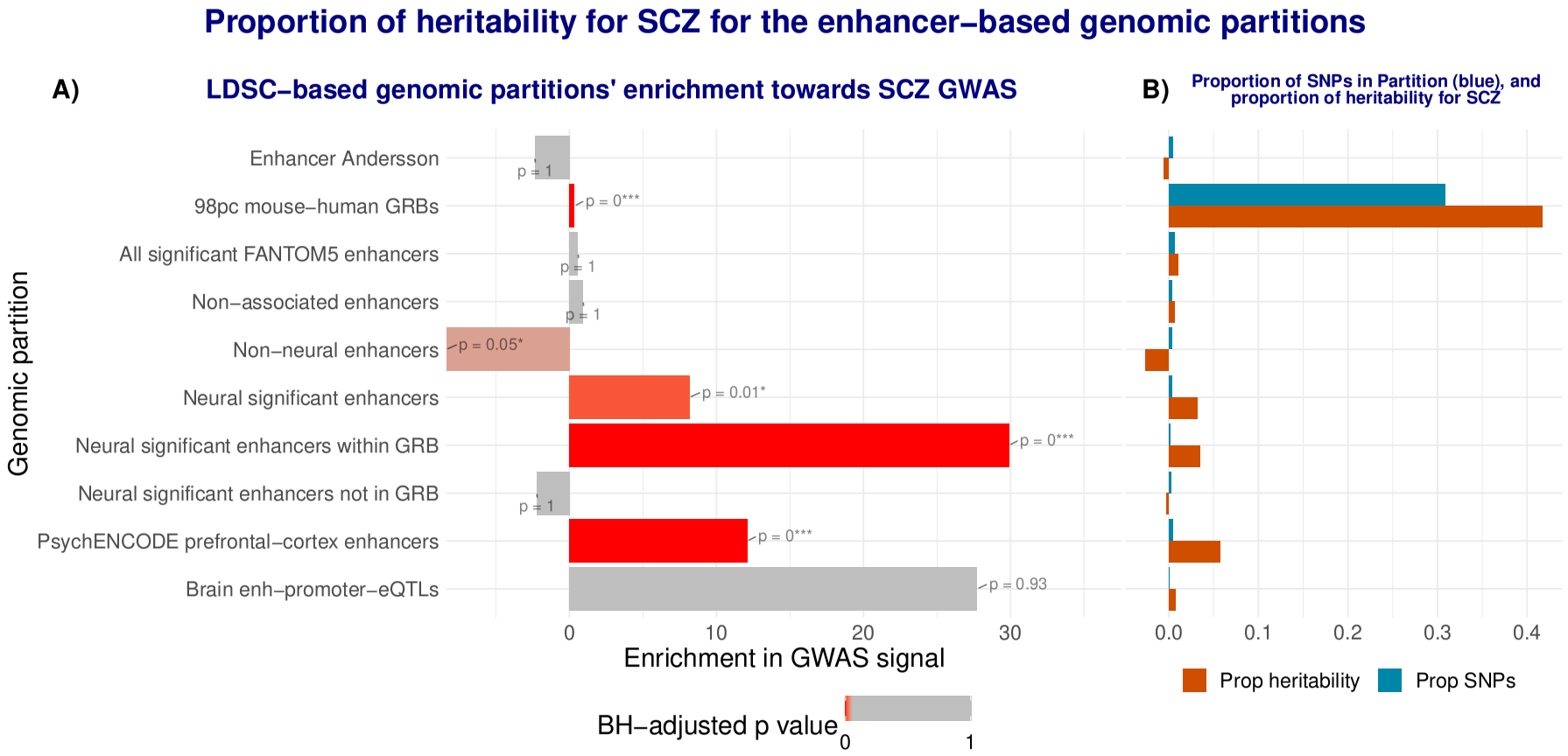


Figure 2.2: Partitioned LDSC for schizophrenia; enhancer-based genomic partitions.
Partitioned LD score regression (LDSC) of several enhancer-based genomic partitions for the latest schizophrenia GWAS (Trubetskoy et al., 2022).
Panel A shows the enrichment value – a positive value indicates more SNPs for schizophrenia in this partition than expected by chance, and a negative value indicates depletion. *P*-values are corrected for multiple testing using the Benjamini-Hochberg (BH) method.
Panel B shows the proportion of SNPs (Prop SNPs in blue) and of schizophrenia heritability for the SNPs in the partition (Prop heritability, in red) for each genomic partition.
Abbreviations: SCZ: schizophrenia; base: all GWAS SNPs; 98pc mouse-human GRBs: GRBs with nearly complete (> 98%) sequence homology between human and mouse; Andersson: (Andersson et al., 2014).

significant AR+C promoter interactions not overlapping GRBs, which were not significantly associated to a schizophrenia signal (BH-adjusted p=1). Finally, the positive control list of ∼18k PsychENCODE pre-frontal cortex specific enhancers was also highly enriched, with a value of 12.1, and BH-adjusted $p< 0.0001$. Enhancers with significant AR+C promoter interactions overlapping significant GTEx brain eQTLs (BRAIN ENH-PROMOTER-EQTL) showed an enrichment value of 27.7, but this was not statistically significant, likely due to the small sample size (BH-adjusted p=0.93).

The enhancer lists that were significantly enriched or depleted in schizophrenia genetic heritability in this analysis – namely the FANTOM5 neural-expressed, as well as the GRB overlap subset, and the negative control lists including the significantly depleted non neural list – will form the basis for the analyses in the next chapters, allowing to generate partitioned polygenic risk scores. However, the PsychENCODE enhancer list was not brought forward for analysis, as no information about tissue-specific expression for this partition was available in FANTOM5.

2.4.2.2 Cardiac-tissue-specific enhancers and HCM heritability

In this section, I have evaluated various genomic partitions to measure enrichment in HCM heritability. I first tested common genomic annotations for enrichment, such as exons, introns, or untranslated regions (3' or 5' UTRs) of genes. Then I tested regions of interest, specifically enhancer-based partitions.

Standard genomic partitions

Once more, I first compared the main standard genomic partitions with a 'base' partition containing all GWAS SNPs with regards to enrichment in HCM heritability. Figure 2.3 shows that neither coding (exons), nor non-coding (introns) gene regions showed significant enrichment in HCM heritability. The same can be said of peri-genic regions, including promoters, promoter flanking, 5' and 3' untranslated regions, which were all not significantly different from the null in terms of HCM heritability. Of the general enhancer

partition that I tested, super enhancers were the largest partition, including 16% of SNPs, and were significantly enriched in HCM heritability, with 37.6% (adjusted $p < 0.001$), while all other categories of enhancers were not significantly enriched or depleted in HCM heritability (Figure 2.3).

Tissue-specific and control enhancer genomic partitions

In this section I tested whether enhancer-based partitions, as well as a human-mouse GRB partition, were enriched in HCM heritability. GRBs hosted a large proportion, 31%, of HCM SNPs, however the partition's heritability was also of 31% (enrichment BH-adjusted p=1). Unlike schizophrenia, here GRBs were therefore not enriched in HCM heritability. To test whether enhancers were generally important for HCM genetics, I then tested enhancer-based partitions. As shown in the previous paragraph, the partition containing all FANTOM5 enhancers (Enhancer Andersson in Figure 2.4; BH-adjusted p=1) was not enriched for HCM heritability; neither were all enhancers with significant AR+C promoter interactions – unselected for tissue expression (ALL SIGNIFICANT FANTOM5 ENHANCERS; BH-adjusted p=1). As expected, FANTOM5 non-associated enhancers (e.g., those not associated to any nearby promoter) were not enriched for HCM heritability either.

When testing tissue-specific enhancer lists, it was evident that FANTOM5 promoters with no heart expression (NON CARDIAC ENHANCERS in Figure 2.4; either significantly associated to a gene or not) did not appear significantly depleted in HCM heritability, with an enrichment value of -16.5, however BH-adjusted p = 0.29. Of the heart-expressed partitions, the largest, counting all ∼9k enhancers with significant AR+C promoter interactions (CARDIAC SIGNIFICANT ENHANCERS) did not appear significantly enriched in HCM heritability, with an enrichment value of 39.4, but BH-adjusted p=0.30. Both the GRB and the non-GRB subsets of this list appeared non-significantly enriched or depleted in HCM heritability. Enhancers with significant AR+C promoter interactions overlapping significant GTEx heart eQTLs (HEART ENH-PROMOTER-EQTLS) showed an enrichment value of 59.1, but this was not statistically significant, likely due to the small sample size (BH-adjusted p

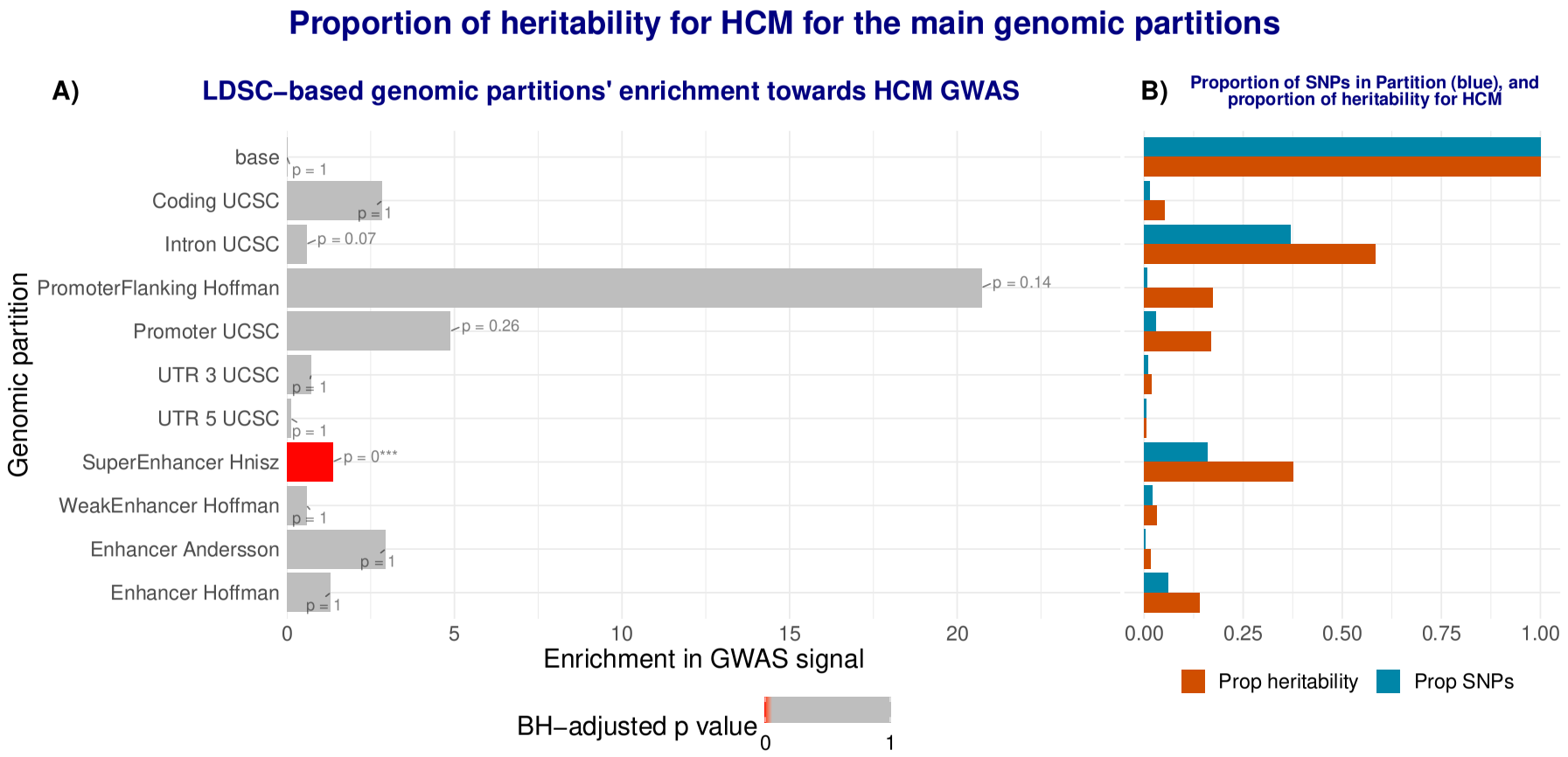


Figure 2.3: Partitioned LDSC for HCM; main genomic partitions.
Partitioned LD score regression (LDSC) of the main genomic partitions for the latest HCM GWAS (Tadros et al., 2023).
Panel A shows the enrichment value – a positive value indicates more SNPs for HCM in this partition than expected by chance, and a negative value indicates depletion. *P*-values are corrected for multiple testing using the Benjamini-Hochberg (BH) method.
Panel B shows the proportion of SNPs (Prop SNPs in blue) and of HCM heritability for the SNPs in the partition (Prop heritability, in red) for each genomic partition.
Abbreviations: Base: all SNPs; UCSC: University of California Santa Cruz database; UTR 3: 3' untranslated gene regions; UTR 5: 5' untranslated gene regions; Hoffman: (Hoffman et al., 2013); Andersson: (Andersson et al., 2014); Hnisz: (Hnisz et al., 2013).

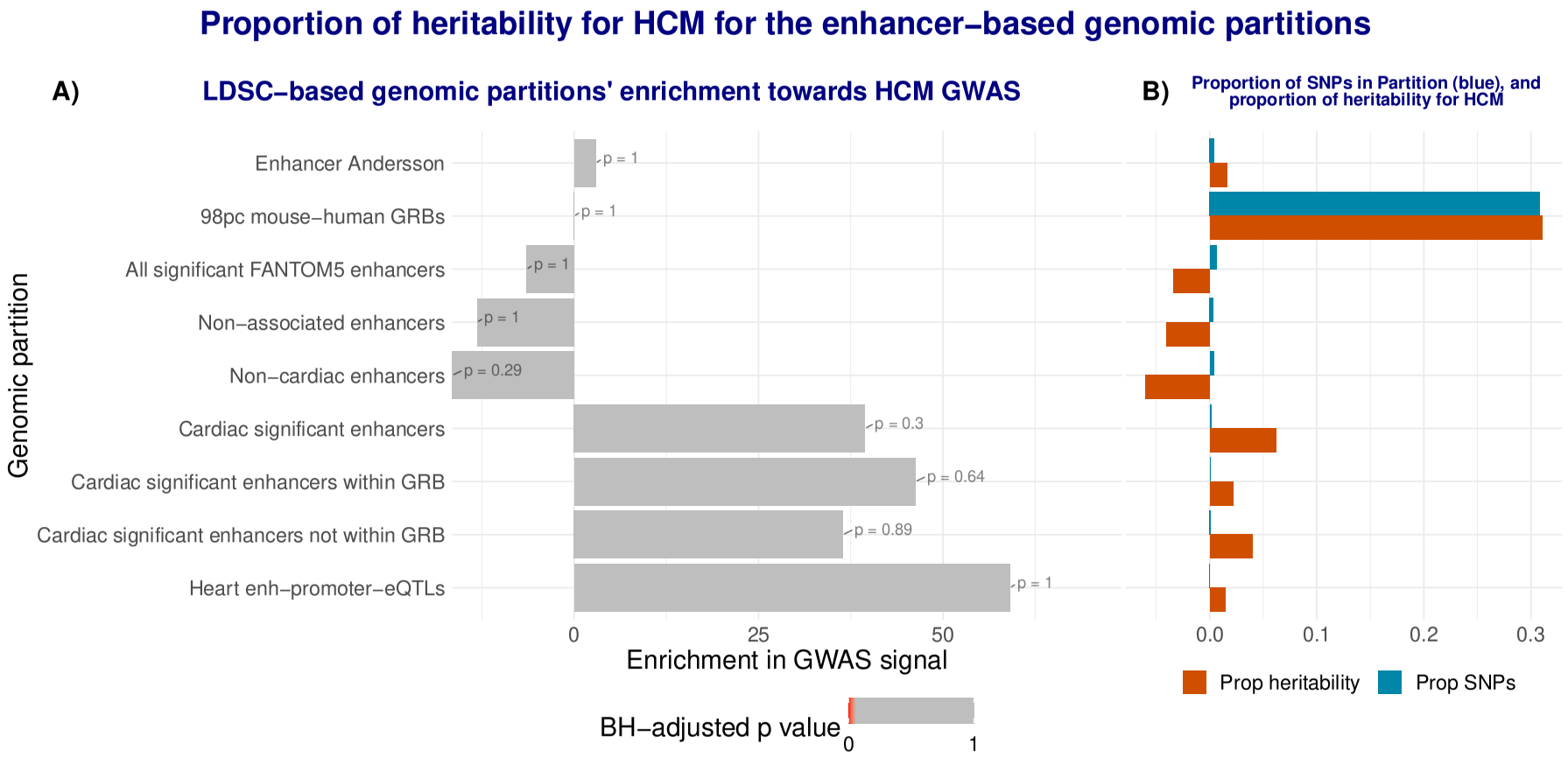


Figure 2.4: Partitioned LDSC for HCM; enhancer-based genomic partitions.
Partitioned LD score regression (LDSC) of several enhancer-based genomic partitions for the latest HCM GWAS (Tadros et al., 2023).
Panel A shows the enrichment value – a positive value indicates more SNPs for HCM in this partition than expected by chance, and a negative value indicates depletion. *P*-values are corrected for multiple testing using the Benjamini-Hochberg (BH) method.
Panel B shows the proportion of SNPs (Prop SNPs in blue) and of HCM heritability for the SNPs in the partition (Prop heritability, in red) for each genomic partition.
Abbreviations: Base: all SNPs; 98pc mouse-human GRBs: GRBs with nearly complete (≥98%) sequence homology between human and mouse; Andersson: (Andersson et al., 2014).

= 1).

2.5 Summary of findings

In this chapter I first created lists of neural tissue-expressed enhancers with significant enhancer-promoter association (co-expression and 3D chromatin contact in the AR+C). I annotated $\sim 21K$ neural-specific enhancers, of which $\sim 8K$ overlapped human-mouse GRBs. I also annotated $\sim 9K$ cardiac-expressed enhancers, of which $\sim 3K$ overlapped human-mouse GRBs. I then compared the enrichment in schizophrenia heritability for a number of genomic partitions, including some standard, generic partitions (e.g., introns or exons), with neural-tissue-specific enhancer partitions, as well as with several other enhancer-based partitions. I then performed the same analysis using cardiac-tissue-specific enhancer partitions, but this time comparing their heritability for HCM. I found that coding genes, introns, super-enhancers (in general), as well human-mouse GRBs and neural-specific enhancers show significant enrichment in schizophrenia heritability. Interestingly, the enrichment in schizophrenia heritability of the neural-specific enhancers partition was explained by GRB overlap: neural-specific enhancers not residing in GRBs did not show enrichment for schizophrenia heritability, while GRB-based neural-specific enhancers showed the strongest enrichment signal. There were much fewer significantly enriched tissue-specific genomic partitions for HCM heritability. These included just super-enhancers, while GRBs and cardiac-expressed enhancers were not significantly enriched for HCM heritability. The findings are discussed in Chapter 5.

Chapter 3

Schizophrenia and HCM heritability from partitioned polygenic risk scores

3.1 Introduction

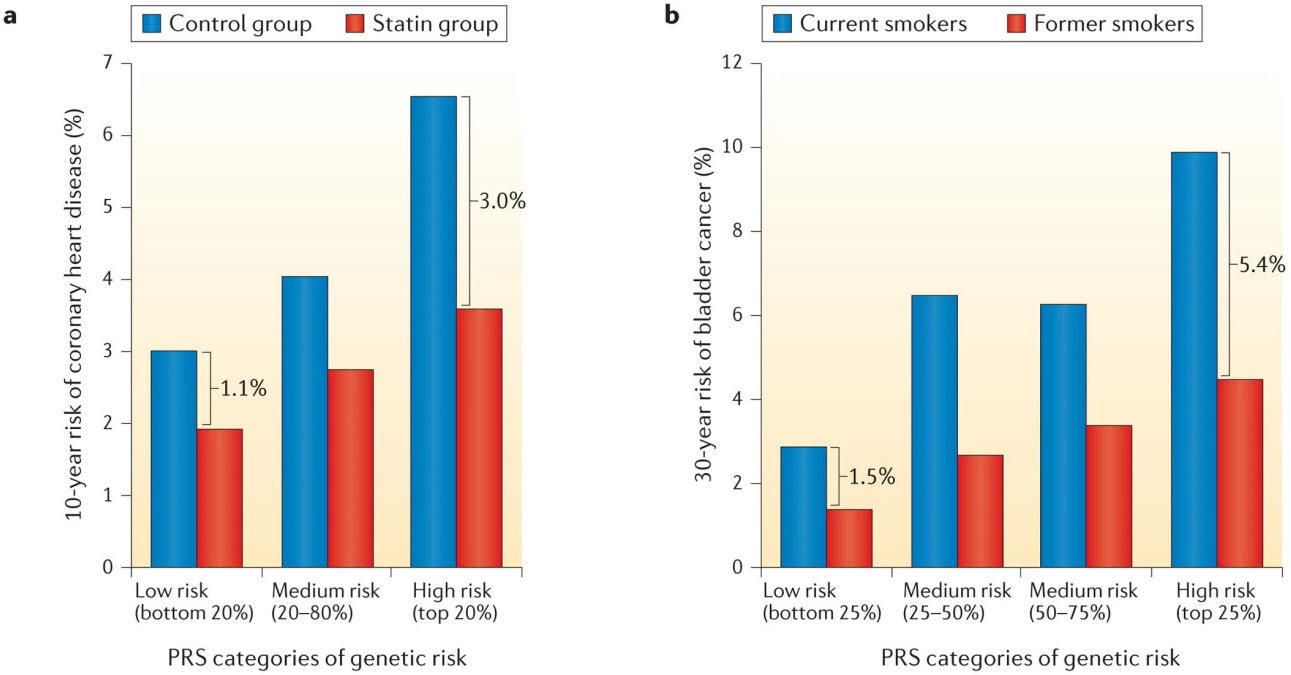
3.1.1 Risk prediction and the contribution of polygenic risk scores

Medicine is a field where it is important to make decisions about a patient's treatment, informed by the likely outcomes of doing nothing, or intervening with one of a multitude of treatments. Increasingly, these decisions are aided by the use of clinical risk prediction models. These are statistical models that aim to predict the probability of future (usually adverse) events, such as a myocardial infarction, by taking into account multiple predictors, e.g., age, sex, ethnicity, and intervening treatments (Grant et al., 2018). Powerful prediction models for important health outcomes exist, and are already embedded in clin-

ical practice: for example, the QRISK score (Hippisley-Cox et al., 2007), now at version 2, allows to calculate the 10-year risk of a person having a cardiovascular event, and its use is already recommended by regulators in the UK for stratifying people for the risk of cardiovascular disease, thus allowing the targeted prescribing of preventative treatments such as statins (National Institute for Health and Care Excellence, 2023).

Calculating polygenic risk scores, as we have seen in the general introduction, in section 1.5, is a powerful technique to express an individual's genetic liability to a condition, based on their genotype. Polygenic risk scores are increasingly being studied as the basis for creating wider risk prediction models for complex disorders, such as cancer (Chatterjee et al., 2016) or cardiovascular disease (Elliott et al., 2020). There are increasing expectations that combining canonical risk predictors, such as age, sex, ethnicity and smoking status, with personalised PRSs, might lead to much more accurate predictions. Figure 3.1 illustrates existing evidence, that shows that subjects at higher polygenic risk may benefit more (that is, have a greater reduction in absolute risk) from risk-reducing interventions, such as statin therapy for preventing cardiovascular disorder, or smoking cessation for preventing bladder cancer. In the example in panel a), in the ASCOT trial, the risk of developing cardiovascular disease was reduced following statin therapy from 3.0% to 1.9% (1.1% absolute risk reduction) among individuals in the lower quintile of genetic risk, whereas it was reduced from 6.6% to 3.6% (3.0% absolute risk reduction) among individuals in the highest quintile of genetic risk – three times the benefit (Chatterjee et al., 2016). This is an example of how, for example, PRSs will be able to help clinicians target interventions to high-risk groups, thus saving money and reducing medication side-effects.

Risk prediction models for several outcomes are being developed in the field of psychiatry, however few are clinically useful, as they either require research-grade information (such as complex scans and tests), or they require information not available at baseline, when the prediction is useful, or else they have not been developed following best practice, e.g., they are not externally validated (Perry et al., 2021a). Some of the few existing models that show clinical promise in the field have the limitation of not being accurate enough, espe-



ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46

ARR (%)	1.5	3.8	2.9	5.4
RRR	0.52	0.60	0.46	0.55

Nature Reviews | Genetics

Figure 3.1: PRS-related risk reduction for coronary heart disease and bladder cancer.
10-year risk of coronary heart disease associated with statin therapy (a) and 30-year risk of bladder cancer associated with smoking status (b), across genetic risk categories defined by PRS. Brackets indicate the absolute risk reduction (ARR) between treatment or exposure groups for subjects in different PRS categories. The tables show the ARR and relative risk reduction (RRR) between treatment or exposure groups (panel a, statin versus control group; panel b, former versus current smokers), across PRS categories.
Figure taken from Chatterjee et al., 2016; Licensed from Springer Nature through RightsLink License Number 5513580975759.

cially for starting higher-risk interventions – or at higher *propensity to intervene thresholds*, to use decision curve analysis jargon. For example, alongside others I have recently developed the *MOZART* algorithm, a risk prediction model for treatment resistance in schizophrenia (Osimo et al., 2023). This algorithm performs well – including in external samples – when predicting lower risks, however it is not useful at higher risk thresholds, due to insufficient sensitivity. It is specifically for overcoming these kinds of limitations – thus hopefully making this and other similar models more useful – that adding PRSs to clinical risk prediction algorithms might make a difference. Adding a further, independent predictor to such models – such as schizophrenia PRS – might add significant amounts to the variance explained by the model, and therefore make them more discriminative.

3.1.2 Tissue-specific enhancers and associations with schizophrenia

We have just seen that PRSs offer promise in terms of personalised risk prediction. However, as we have seen in section 1.5.2, current PRSs only explain a fraction of the genetic liability to most serious mental illnesses. In the case of schizophrenia, the latest and largest GWAS to date explains up to $\frac{1}{3}^{rd}$ of the genetic liability for the condition (Trubetskoy et al., 2022). This is why, in this chapter, I am aiming to increase the proportion of variance explained by the genetic factor for schizophrenia, by taking into account functional annotations regarding enhancers, which I (in chapter 2) and others (e.g., Barešić et al., 2020) have shown to be hosting much more than the expected heritability for schizophrenia.

Furthermore, I am aiming to use information about enhancer **tissue-specific expression**. As we have seen in the general introduction and in paragraph 2.1.1, the non-coding genome plays a crucial role in gene expression regulation through enhancer-promoter interactions, which appear to be both tissue- and time-specific. In particular, GRBs, which are characterised by a target-bystander gene structure, have been shown to host enhancers that play a particularly important role in regulating the expression of developmentally relevant target genes (Akalin et al., 2009). Further, Georgieva, 2022 has found that predicted GRB target genes were enriched in gene ontologies including *axon development*, *embryonic*

organ development, *forebrain development*, *axon guidance*, and *neuron projection*, among others. It is also known that schizophrenia is a neurodevelopmental condition, with genetic and environmental signals converging on the synapse, as discussed in paragraphs 1.6.1 and 5.1.1 and in Howes and Onwordi, 2023.

For all these reasons, I have hypothesised that the h^2_{pPRS} for schizophrenia and HCM will increase considering *tissue-specific enhancers* and *residual* partitions separately, as compared to the original GWASes for the conditions. To test this hypothesis, in this chapter I will use the latest GWAS summary statistics for schizophrenia (from Trubetskoy et al., 2022) and for HCM (from Tadros et al., 2023) as base GWASes (see section 1.5.1), and I will calculate PRSs and partitioned PRSs (pPRSs, see section 3.3.1.2 below) on PGC Consortium cohorts (see paragraph 3.2.1) as target populations for schizophrenia, and on the UK Biobank (see paragraph 3.2.3) and on the Royal Brompton HCM cohort for HCM. In brief, 'partitioned' polygenic risk scores are PRSs where two genomic partitions (e.g., **prioritised** *tissue-specific enhancers* and *residual* partitions) are considered separately for polygenic risk scoring, and then disease heritability is calculated separately for the original GWAS for the condition (h^2_{SNP}), as well as for the partitioned PRSs (h^2_{pPRS}). I will also test if *tissue-specific enhancer*-based PRSs for schizophrenia and HCM will increase the overall disease h^2_{pPRS} when accounting for tissue-specific enhancer expression or target gene association measures – by multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by either the *effect size* of the tissue-specific enhancer, or by its tissue-specific expression. All of this is explained in more detail in the next few paragraphs.

3.2 Materials

3.2.1 Wave 3 Psychiatric Genomic Consortium Schizophrenia population and GWAS

The core Psychiatric Genomic Consortium (PGC) wave 3 schizophrenia dataset comprises 90 international cohorts, including individual-level genotype data (Trubetskoy et al., 2022). This core dataset contains genotypes on 161,405 unrelated individuals; 67,390 cases of schizophrenia or schizoaffective disorder, and 94,015 control individuals. Around 80% of the participants (53,386 cases and 77,258 controls) are of European ancestry, and only these were included in the present analysis.

For this work I used three of the five largest PGC cohorts and the relative leave-one-out genome-wide association studies for schizophrenia as base GWASes to avoid any overlaps between development and validation samples. In particular, I included the *clz2a* cohort, the largest in Trubetskoy et al., 2022, with over 5000 participants with treatment-resistant schizophrenia; the *celso* cohort, the third largest with over 2000 people with schizophrenia recruited across eight Spanish inpatient psychiatric units; and the *xs234* cohort, the fifth largest cohort with over 2000 people with schizophrenia from the Swedish population. The *xclm2* and *xclo3*, the second and fourth largest cohorts, were not included, as they represented the same geographical and clinical category as *clz2a*, i.e., UK patients with treatment resistant schizophrenia. Results for the *xs234* cohort are the main schizophrenia results, as the patients in this cohort – from a population-based Swedish sample – appeared to be the most representative of a general, White ethnic population, while results for the *clz2a* and *celso* cohorts are presented as a sensitivity analysis in the Appendices.

3.2.2 Hypertrophic cardiomyopathy case-control GWAS and Royal Brompton Hospital target population

The HCM GWAS I used for this work is the largest to date, provided to me ahead of publication by Tadros et al., 2023. This includes HCM cases and controls from 7 strata: the Hypertrophic Cardiomyopathy Registry (HCMR), a Canadian HCM cohort, a Netherlands HCM cohort, the Genomics England 100K Genome Project (GEL), the Royal Brompton HCM cohort, an Italian HCM cohort and the BioResource for Rare Disease (BRRD) project. The sample included 5,900 HCM cases, 68,3593 controls, and 36,083 UK Biobank (UKB) participants with cardiac magnetic resonance imaging available to rule out HCM. To be described as cases, participants had to show unexplained left ventricular hypertrophy. This was defined as a left ventricular wall thickness >15mm, or >13mm and either presence of family history of HCM, or a pathogenic or likely pathogenic genetic variant for HCM. Further details can be found in Tadros et al., 2023.

For this work I also used the unpublished leave-one-out GWAS excluding the Royal Brompton HCM cohort (one of seven included cohorts) from Tadros et al., 2023. The same Royal Brompton HCM cohort genotypes were the matched target population. This cohort included 448 patients with HCM and 1219 matched healthy controls. Cases were unrelated British HCM patients from the Royal Brompton & Harefield Hospitals NHS Trust Cardiovascular Research Biobank. For further details, please see (Tadros et al., 2023).

3.2.3 UK Biobank

The UK Biobank project is a large-scale study that collected deep genetic and phenotypic data on ∼500,000 individuals aged between 40 and 69 from across the United Kingdom. This resource provides a variety of phenotypic and health-related information on each participant, including biological measurements, lifestyle indicators, biomarkers, and imaging, as well as mapping to each participant's NHS records (which provide, for example, diagnostic information). All participants provided DNA for genotyping (Sudlow

et al., 2015).

As detailed in Bycroft et al., 2018, genotype calling was performed by Affymetrix on two purpose-designed arrays. ∼50,000 participants were run on the *UK BiLEVE Axiom array*, and the remaining ∼450,000 were run on the *UK Biobank Axiom array*. There are 805,426 markers in the released genotype data. The genotype data were quality controlled (QC). In addition, the dataset was phased and ∼96 million genotypes were imputed using computationally efficient methods combined with the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) and UK10K haplotype (https://www.sanger.ac.uk/collaboration/uk10k/) resources.

Given that target data should mirror the ethnic mix of base data, I excluded participants of non White ethnicity and with a kinship coefficient $> 12.5\%$. UKBB includes 1,605 cases of schizophrenia, and 597 cases of HCM.

3.3 Methods

As discussed in section 1.5.1, PRSs are generated by applying the effect sizes from a GWAS (in PRS terms, the **base dataset** – a GWAS summary) onto a genotyped cohort (in PRS terms, the **target dataset**); the sample from which the GWAS was generated and the target cohort should ideally not overlap.

For this work, the target populations for schizophrenia are three of the five largest PGC wave 3 cohorts (Trubetskoy et al., 2022), as described in section 3.2.1 above. For each cohort, the relative leave-one-out (LOO) GWAS was used as the base dataset. I present the results on the *xs234* cohort as the main schizophrenia results, while results for the *clz2a* and *celso* cohorts are presented as sensitivity analyses in the Appendix. The base dataset for the HCM GWAS is the meta-analysis described in paragraph 3.2.2, as generated in (Tadros et al., 2023). The target dataset is the UK Biobank (Sudlow et al., 2015). A sensitivity analysis was conducted on the leave-one-out GWAS excluding the Royal Brompton HCM cohort (one of seven included cohorts) from Tadros et al., 2023, and the same Royal Brompton HCM cohort

genotypes as the matched target population.

As part of work for this thesis I have generated tissue-specific enhancer lists (Chapter 2). These enhancers are likely to be significant in regulating genes in brain (for the neural-specific lists) and in heart (for the cardiac-specific lists), as they were selected on the basis of significant co-expression with promoters, and 3D-genome interactions, as part of the AR+C (Georgieva, 2022), as well as on the basis of their expression in the tissue in question.

Figure 3.2 summarises the steps involved in generating partitioned PRSs (pPRSs), and in comparing them with the original GWAS PRS. In brief, for each base-target paired dataset, and for each tissue-specific enhancer list:

1. Three 'base datasets' are used. The original GWAS (or leave-one-out GWAS), the *enhancer-based* and the *residual* subsets of the original GWAS.

2. The original GWAS and each of the subsets are QCed and clumped. Of note, the *enhancer-based* and *residual* partitions are clumped together, so that there is no LD overlap between the SNPs; in this process, SNPs within the enhancer partition are prioritised over those in the residual partition – e.g., if there are two SNPs in the same LD block, one falling within an enhancer, and one falling outside of it, the enhancer-based SNP is retained even if it has a higher disease association *p*-value, due to the *a-priori* hypotheses about the importance of long-range and tissue-specific gene regulation by enhancers in schizophrenia (see sections 2.1.1 and 3.1.2).

3. A disease PRS is calculated for each of the three partitions on the target population (see next section for the technical detail about how the calculation is performed).

4. Modelling is performed to calculate the amount of heritability explained by each base dataset, on its own and combined (e.g., combining the *enhancer-based* and *residual* partitions) – see section 3.3.2.

5. Finally, two more modified PRSs are calculated and modelled onto the target population, by multiplying SNP-disease association measure β coefficient by either the *effect*

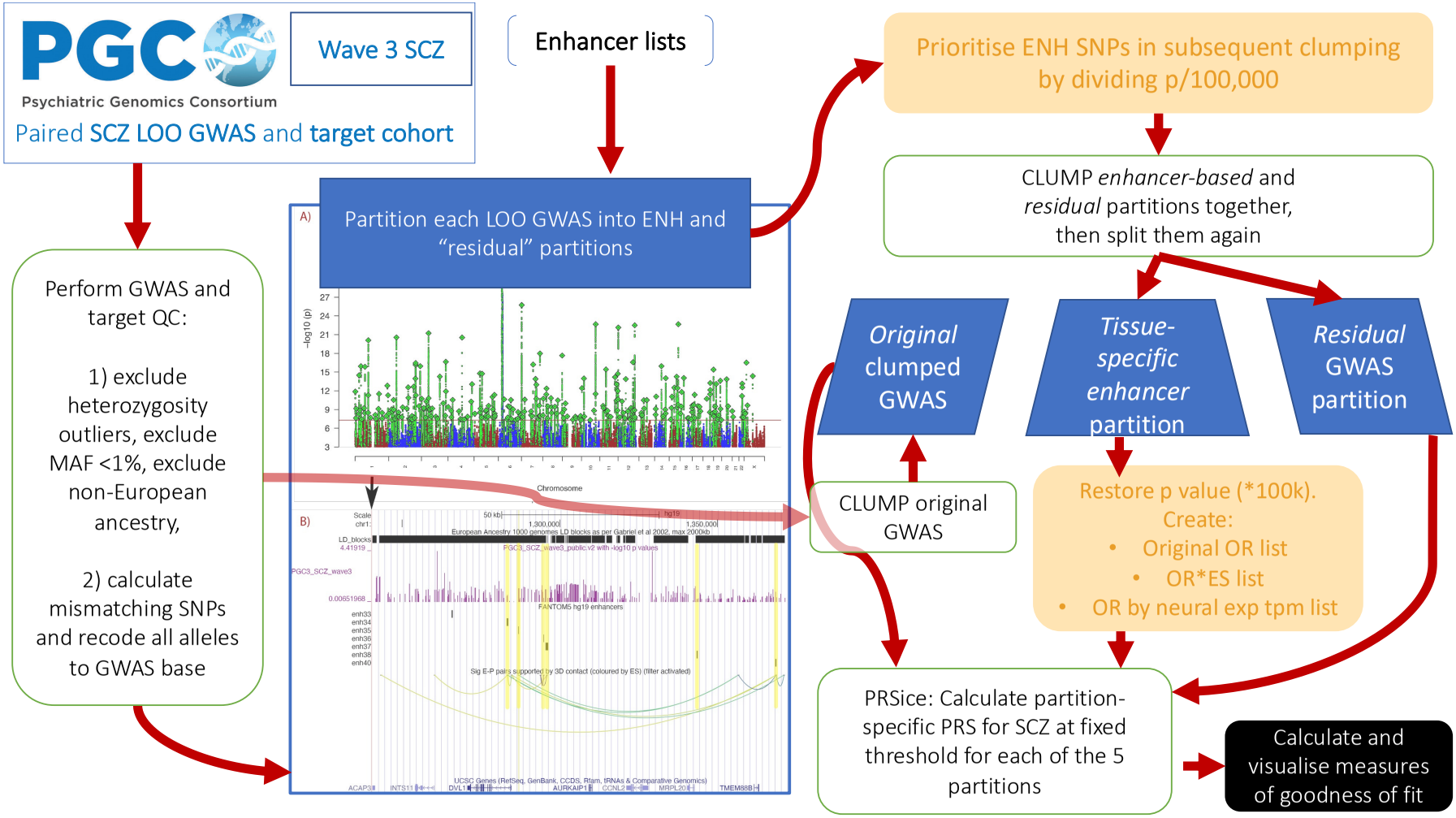


Figure 3.2: Graphical methods for 'partitioned' PRS generation for schizophrenia.
The figure describes the methods for this chapter. Two sets of PRSs are generated in parallel: a canonical PRS, generated using the C+T method, on the base GWAS is calculated, as well as PRSs for each of two genomic partitions: an *enhancer-based* and a *residual*. The amount of variance explained by each is then compared.
Abbreviations: SCZ: schizophrenia; LOO: leave-one-out; ENH: enhancer; QC: quality control; GWAS: genome-wide association study; PGC: Psychiatric Genomic Consortium; p=p value; OR: odds ratio.

size of each tissue-specific enhancer (its association measure with a target gene), or by its tissue-specific expression.

I will go into specific details for each step over the next sections.

3.3.1 PRS calculation

3.3.1.1 GWAS (base) and target quality control (QC)

The base datasets – the GWASes used to build the PRS – are first quality controlled (QCed) by removing rare variants (MAF < 0.01) and those with INFO score < 0.8. The target dataset – the cohort on which the PRS will be calculated – is then quality controlled. This entails:

- Standard target QC, including filtering out variants with a MAF < 0.01, INFO score < 0.8, SNPs in controls with $p < 10^{-15}$ in a Hardy-Weinberg Equilibrium Fisher's exact test (which are more likely affected by genotyping error or the effects of natural selection); SNPs that are missing in a high fraction of subjects ($> 10\%$), since this may indicate problems in the DNA sample or processing.
- Target pruning to remove highly correlated SNPs; then, on pruned SNPs, removing individuals with high heterogeneity – measured as an F coefficient > 3 standard deviations (SD) from the mean – as this might indicate DNA contamination or high levels of inbreeding.

Then, considering the specific base and target datasets, the two datasets are harmonised by:

- Strand flipping – e.g., recoding SNPs encoded as A>T in the base and T>A in the target dataset.
- Removing mismatching SNPs – e.g., SNPs for the same position where the alternative allele did not match, such as A>T in the base and A>G in the target dataset.

3.3.1.2 GWAS (base) partitioning and clumping

The *base* GWAS is then divided into paired *enhancer-based* and *residual* partitions for each enhancer list, by separating SNPs falling inside enhancers ($\pm 100bps$). The two lists then need clumping together, so that a global PRS can be calculated without the risk of including multiple SNPs per LD block. Clumping, as discussed in paragraph 1.4.3.1, works by selecting the top variant (i.e., the SNP with the lowest *p*-value) for each clump. Therefore, to prioritise enhancer-residing SNPs, enhancer-based SNP *p*-values are divided by 100,000 before clumping, so that they are retained with priority over nearby residual SNPs. The prioritised enhancer SNP list and the residual list are then clumped together to only retain one SNP per LD block per paired *enhancer-based* and *residual* partitions for each enhancer list.

The settings for clumping both the LOO GWAS, as well as the paired *enhancer-based* and *residual* partitions, are the PLINK flags:

$--clump{-}p1\ 1\ --clump{-}p2\ 1\ --clump{-}kb\ 500\ --clump{-}r2\ 0.1$.

This means that only one SNP is retained for each 500Kb window, or for any block with an LD $r^2 > 0.1$.

3.3.1.3 Annotation and up-weighting of enhancer-based GWAS (base) partition SNP odds ratios

After clumping together, the enhancer-based and residual GWAS (base) partitions are again separated. The enhancer-based SNP *p*-values are multiplied by 100,000 to restore the original *p*-value for the remainder of the analysis.

The enhancer-based GWAS (base) partition is annotated using data from the AR+C, including the maximum effect size (ES) for each enhancer, and the mean log-transformed tissue-expression value for neural or cardiac enhancers.

Three separate measures of SNP-disease association are collected to produce enhancer-based annotated partitions:

- The original GWAS' SNP association odds ratio (OR), which is the canonical measure to use in PRS scoring.
- A product of the SNP association OR and the AR+C *effect size (ES)* for the specific enhancer – but transformed to obtain a value comparable in scale to the original OR. Keeping in mind that $\beta = ln(OR)$, this is calculated as:

$$OR_{ES} = e^{(\beta \times ES)} \tag{3.1}$$

- A product of the SNP association OR and the enhancer tissue-specific expression (TS_{exp}) in transcripts per million (tpm) for the specific enhancer, calculated as:

$$OR_{TS_{exp}} = e^{(\beta \times TS_{exp})} \tag{3.2}$$

Partitioned PRSs are then calculated for each of these three measures for each *enhancer-based* partition, for each enhancer list.

3.3.1.4 PRS calculation

PRSs are calculated using PRSice 2.3.5 (Choi & O'Reilly, 2019; Euesden et al., 2015) using the clumping and thresholding method (see section 1.5), and selecting a fixed and pre-defined *p*-value threshold of 0.5 (and 0.05 as a sensitivity analysis in the Appendix), so that PRSs are comparable across cohorts and lists.

As variants have already been clumped in one of the steps above, PRSice is run with the flags:

$--no\text{-}clump\ --keep\text{-}ambig\ --quantile\ 3\ --binary\text{-}target\ T\ --prevalence\ X$

with prevalences of 0.01 for schizophrenia (Saha et al., 2005), and of 0.00225 for HCM (Marian & Braunwald, 2017). The covariates included age, sex, and the first 10 PCA components of the genotypes for the target population.

3.3.2 Calculation of the proportion of variance explained by the genetic factor – or coefficient of determination

As described above, I produced a PRS for each genomic partition (original clumped GWAS, tissue-specific *enhancer-based* and *residual* partitions), for each enhancer list (e.g., NEURAL SIGNIFICANT ENHANCERS), for each target population/base GWAS pair.

To calculate the proportion of variance explained by the genetic factor for each target population I used a logistic model in the form of: $dx \sim PRS$, with a logit link function to obtain measures of model fit – summarising the amount of the diagnostic variance that is explained by the PRS. This also allowed to calculate standard measures of model fit, including a Nagelkerke pseudo-R^2. The Nagelkerke pseudo-R^2 is however liable to over or under-estimations of the variance explained, as suggested in Lee et al., 2012, depending on the degree of genetic disease liability, and on the target population's prevalence of disease. This is often high in case-control samples (usually around 50%), which can lead to over-estimations of the variance explained.

To convert the raw Nagelkerke pseudo-R^2 onto the liability scale for the condition, and adjust for ascertainment bias, I converted the raw performance measures of such a model into the proportion of the total variance explained by the genetic factor on the liability scale, and corrected this for ascertainment, as per Lee et al., 2012.

$R^2_{O_{CC}}$, is the proportion of the total variance explained by the genetic factor on the observed probability scale for an ascertained case-control study:

$$R^2_{O_{CC}} = \left[z \times \frac{\sqrt{P \times (1-P)}}{K \times (1-K)} \right]^2 \times \frac{\sigma^2_g}{\sigma^2_{gCC}} \times \sigma^2_g \tag{3.3}$$

Where:

- K is the population disease prevalence;
- P is disease prevalence in the case-control sample;
- z is the height of a normal density curve at the point that truncates the proportion K

in the upper tail;

- g_{CC} is the genetic liability for the case-control (CC) condition.

From this formula one can derive $R^2_{l_{CC}}$, or the proportion of the total variance explained by the genetic factor on the liability scale, corrected for ascertainment:

$$R^2_{l_{CC}} = \sigma^2_g = \frac{R^2_{O_{CC}} \times C}{1 + R^2_{O_{CC}} \times \theta \times C} \tag{3.4}$$

Where:

- m is the mean liability for cases
- $C = \frac{K \times (1-K)}{z^2} \times \frac{K \times (1-K)}{P \times (1-P)}$
- $\theta = m \times \frac{P-K}{1-K} \times \left(m \times \frac{P-K}{1-K} - t \right)$
- t is the threshold on the normal distribution truncating the proportion of disease prevalence K, and θ

For this work, a modified version of this formula was used, derived by Sam Choi for Choi and O'Reilly, 2019, as documented at https://groups.google.com/g/PRSice/c/kqLKYUhHfhM/m/EMck3Pa9BQAJ:

$$R^2_{Choi} = \frac{R^2_{Nagel} \times C \times e}{1 + R^2_{Nagel} \times C \times e \times \theta} \tag{3.5}$$

Where:

- R^2_{Nagel} is the Nagelkerke pseudo-R^2 from the logistic model
- $e = P^{2 \times P} \times (1-P)^{2 \times (1-P)}$

This allowed to a) adjust the output for total disease liability, and b) adjust the output – obtained from a case-control dataset where about 50% of the participants had the disease in PGC cohorts, or where a very small minority of the population had the condition at hand when using UKBB – to comparable values.

3.3.3 Software

Statistical analyses were performed in R 4.2.2 (R Core Team, 2023), using Tidyverse libraries (Wickham et al., 2019). Some genomic manipulations required the use of BioConductor 3.16 (Gentleman et al., 2004) packages including biomaRt (Durinck et al., 2009), GenomicRanges (Lawrence et al., 2013), and rtracklayer (Lawrence et al., 2009). Some plots use colour from the MetBrewer package (Mills, 2022).

Most genomic manipulations and genotype-phenotype associations were performed using PLINK 1.9 and 2 (Chang et al., 2015; Purcell & Chang, 2022). C+T polygenic risk scoring was performed using PRSice 2.3.5 (Choi & O'Reilly, 2019).

A pipeline was built automating most of these steps utilising Nextflow 22.10 (Di Tommaso et al., 2017), using Docker containers hosted on Docker Hub https://hub.docker.com: container emosyne/prsice_gwama_exec:1.0 for PRSice, container emosyne/plink2:1.23 for PLINK and PLINK2, container emosyne/r_docker:1.97 for R and all related packages.

3.3.3.1 Code availability

All code for this work is available on the GitHub repositories:

- Schizophrenia/PGC pipeline: https://github.com/emosyne/lisa_percohort_devel_pub
- HCM pipeline: https://github.com/emosyne/HCM_cardiac_enhs

3.4 Results

In this chapter I will develop 'partitioned' polygenic risk scores, or PRSs where two genomic compartments (e.g., the *tissue-specific enhancers* and the *residual* compartments) are considered separately for polygenic risk scoring, and then disease heritability is calculated separately for the original GWAS for the condition (h^2_{SNP}), as well as for the partitioned PRSs (h^2_{pPRS}). I will also test if multiplying SNP-disease association β for enhancer-based SNPs by either an *effect size* of the tissue-specific enhancer, or by its neural expression, can increase disease h^2_{pPRS}.

To do so, as we have seen in more detail in the Methods (see Figure 3.2), this chapter consists of analyses which start from a base GWAS, a target population, and a list of genomic coordinates, forming an enhancer-based genomic partition (these were generated in Chapter 2). Then, for each base (GWAS), target (population) and enhancer list combination, the base and target datasets are QCed, and three genomic partitions are created: a whole-GWAS partition (called *original GWAS*), and its two subsets, the *tissue-specific enhancer* partition (+100bps at each side), and the *residual* partition. The three partitions are then clumped (the original GWAS on its own, and the *tissue-specific enhancer* and *residual* partitions together). Partitioned PRSs for the condition – schizophrenia or HCM – are then computed for each of the three genomic partitions – using a fixed C+T threshold of 0.5. For the *tissue-specific enhancer* partition, in addition to calculating the standard PRS (which makes use of OR and *p*-value from the base GWAS), I also calculated modified versions of the score, utilising OR multipliers derived from AR+C-derived enhancer tissue-specific expression data. The resulting PRSs are:

- The original clumped GWAS PRS, including all SNPs (after clumping) and using the GWAS-derived original ORs (for comparison).
- Three *tissue-specific enhancer* (TS_ENH) partitions, resulting from all SNPs within 100bps of the specific enhancer list being considered. This partition is calculated as three distinct 'versions': using the original OR for each SNP; using a modified OR enhanced using the AR+C *effect size* measure (ES); using a modified OR enhanced using tissue-specific enhancer expression values.
- A *residual* partition, equivalent to the original base, after subtraction of *tissue-specific enhancer* SNPs, and clumping the two lists together – but prioritising enhancer SNPs.

The output of the analysis is presented as a number of plots, allowing the reader to compare the total variance explained by the genetic factor on the liability scale, corrected for ascertainment – called the coefficient of determination – for each partition, and for models including more than one partition. Please see Figure 3.2 for a graphical summary of the methods for this chapter.

3.4.1 Neural tissue-specific enhancers and variance explained in schizophrenia

Analyses of schizophrenia and neural tissue-specific enhancers were performed on PGC Consortium cohorts. Here I present the results regarding the *xs234* PGC cohort as target, and the relative leave-one-out genome-wide association study for schizophrenia (LOO GWAS) as base GWAS, to avoid overlap between the base and target samples. Results for a further two PGC cohorts (*clz2a* and *celso*) can be found for comparison in the Appendix, section A.1. Additionally, results obtained running the pipeline at the additional threshold of $p = 0.05$ (instead of the main $p = 0.5$) for the *xs234* PGC cohort can be found in Appendix section A.2.

The enhancer lists I tested – those significantly enriched in GWAS signals for schizophrenia in Chapter 2 – were:

1. NEURAL SIGNIFICANT ENHANCERS: $\sim 21K$ neural-expressed enhancers, with significant co-expression with at least one promoter, and with evidence of significant E-P 3D contact.
2. NEURAL SIGNIFICANT ENHANCERS WITHIN A GRB: $\sim 8K$ neural-expressed enhancers, subset from 1), additionally overlapping a human-mouse GRB.
3. NON-NEURAL ENHANCERS: $\sim 20K$ enhancers with no neural expression – not necessarily in a significant enhancer-promoter pair (negative control 1).
4. NON-ASSOCIATED ENHANCERS: $\sim 34K$ FANTOM5 enhancers not associated to a gene – not necessarily with any neural expression (negative/neutral control 2).

3.4.1.1 Patient and SNP selection for the *xs234* PGC schizophrenia target cohort

The *xs234* PGC European ancestry cohort included 2,077 people with schizophrenia, and 2,341 controls. This is one of six batches of the Swedish Schizophrenia Study, collected in a multi-year project; cases were identified from the Swedish population via the Swedish Hospital Discharge Register, which captures all hospitalizations. Controls were selected at

random from Swedish population registers. For more details on PGC samples, consult the latest PGC schizophrenia GWAS paper (Trubetskoy et al., 2022).

The original target *xs234* European population files contained genetic information about 4,418 people, and on 8,957,878 imputed SNPs for each person. Following QC, 4,335 participants remained, and information about 6,758,413 imputed SNPs. The original *xs234* leave-one-out PGC schizophrenia base GWAS based on European populations contained information about 7,660,660 SNPs, based on 51,419 cases and 74,993 controls. After base GWAS QC, information remained about 7,113,397 SNPs, which further reduced to 256,598 following clumping.

In terms of tissue-specific lists, when splitting the base GWAS into an *enhancer-based* and *residual* partitions, the total sum of SNPs in these partitions did not equal to the QCed, clumped original GWAS partition, as the splitting process – as detailed in the methods – was doped by prioritising SNPs in enhancer partitions by temporarily reducing their *p*-value before joint clumping. As an example of how the number of included SNPs changed over the process, I am presenting here data for one such lists, the NEURAL SIGNIFICANT ENHANCERS, in this cohort:

- Number of schizophrenia GWAS SNPs in the NEURAL SIGNIFICANT ENHANCERS partition before clumping: 25,642.
- SNP number in the corresponding *residual* partition before clumping: 7,087,756. The sum of the two before clumping still equated to 7,113,397 SNPs, which is the number of SNPs in the original GWAS partition before clumping.
- After clumping, there were 13,044 SNPs left in the NEURAL SIGNIFICANT ENHANCERS partition, preserving $\sim 51\%$ of pre-clumping SNPs.
- After clumping, there were 242,322 SNPs left in the corresponding *residual* partition, preserving $\sim 3\%$ of pre-clumping SNPs.

Similar results were obtained for all lists and all base GWASes. The exact numbers of SNPs in each partition are presented in each section below.

3.4.1.2 Coefficients of determination for the main genomic partitions in schizophrenia

First, I examined how much of the variance for the schizophrenia phenotype could be explained by the three main genomic partitions: the *original*, clumped GWAS, for comparison; the *enhancer-based* partitions; and the *residual* partitions for each enhancer list. As explained in more detail in the methods section, I focus on comparing the coefficients of determination (CoD) – or the proportion of the total variance explained by the genetic factor on the liability scale, corrected for ascertainment, as in Lee et al., 2012, utilising the formulas by Choi and O'Reilly, 2019. In the figures, I present both the raw Nagelkerke pseudo-R^2 for each logistic model, and the CoD, for comparison. The raw Nagelkerke pseudo-R^2 are not commented on in text. The main analysis focussed on the *xs234* European population.

I found that the original leave-one-out GWAS CoD for schizophrenia in this sample, based on $\sim 161K$ SNPs, equated to 9.85% (95% confidence interval (CI): 8.72; 11.01). The CoD per SNP equated to 6.13×10^{-7} (Figure 3.3). These values of course did not differ between enhancer lists.

With regards to the *enhancer-based* partitions, these had CoDs of 3.72% (95% CI 3.00; 4.45) and of 2.22% (95% CI 1.66; 2.79) for the NEURAL SIGNIFICANT ($\sim 8K$ SNPs), and NEURAL SIGNIFICANT WITHIN GRBS ($\sim 3K$ SNPs) partitions, respectively. The CoD per SNP equated to 47.8 and 77.9 $\times 10^{-7}$, respectively, which equates to between $\sim$8 and $\sim$13 times the value per SNP of the original GWAS (Figures 3.3A and 3.3B).

The control lists (NON NEURAL and NON ASSOCIATED ENHANCERS) showed lower explained variance per SNP as compared to the NEURAL SIGNIFICANT ENHANCERS partitions, with CoDs per SNP of $\sim$26 and $\sim$30 $\times 10^{-7}$, respectively. These values were still between 4 and 5 times the value per SNP of the original GWAS (Figures 3.3C and 3.3D).

Sensitivity analysis in *xs234* at the *p*-value threshold of 0.05

In a sensitivity analysis, I then tested whether comparing PRSs between partitions while using a different threshold of 0.05 instead of 0.5 made a difference in the results. The figures for each enhancer list for this analysis are Appendix Figures A.9 to A.12.

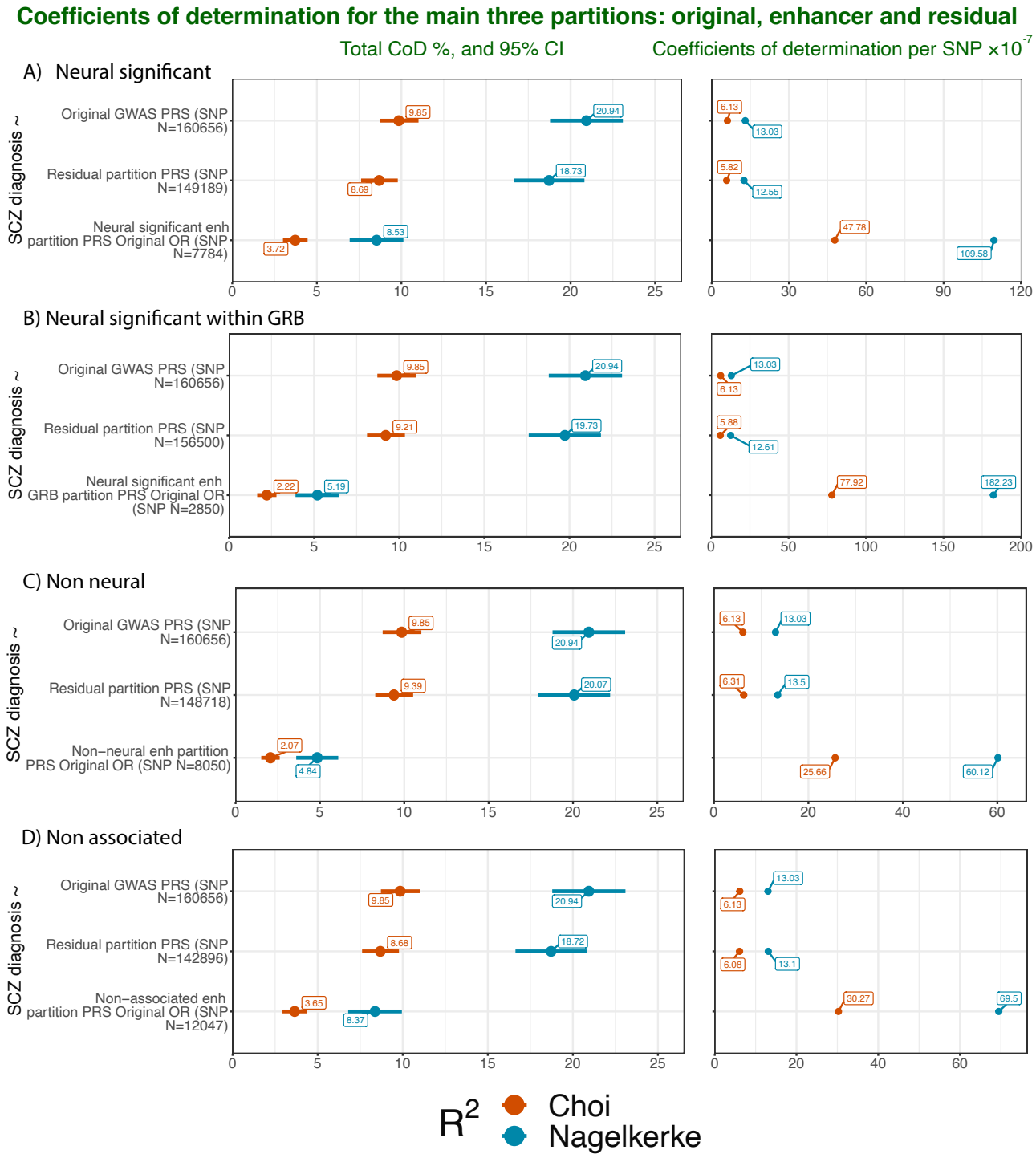


Figure 3.3: Coefficients of determination for schizophrenia for the three main partitions, original GWAS, *residual* and *tissue-specific* enhancers, in the *xs234* cohort.
The figure describes the proportion of the variance of schizophrenia explained by the genetic factor for each PRS for the three main genomic partitions – original GWAS, *residual* and *tissue-specific* enhancers, in the *xs234* cohort. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Each plot on the left shows the overall CoD % and 95% confidence interval, and on the right the corresponding point value adjusted per SNP ($\times 10^{-7}$). **Panels A) to D)** show the CoDs for each genomic partition for the NEURAL SIGNIFICANT, NEURAL SIGNIFICANT WITHIN GRBS, NON-NEURAL, and for the NON-ASSOCIATED lists, respectively.
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

Despite the smaller number of SNPs included at this lower threshold, ∼45k, the leave-one-out GWAS coefficient of determination (CoD) for schizophrenia in this sample equated to 10.28% (95% CI 9.13; 11.47), higher than that at the original threshold of 0.5. The CoD per SNP equated to 22.91 $\times 10^{-7}$ (Appendix figure A.9A).

At this threshold, the NEURAL SIGNIFICANT (∼2k SNPs) and NEURAL SIGNIFICANT WITHIN GRBS (∼800 SNPs) partitions had CoDs of 3.13% (95% CI 2.47; 3.81) and of 1.84% (95% CI 1.32; 2.36), respectively (Appendix figures A.9A and A.10A). These partitions showed CoDs per SNP that were ∼7 and ∼10 times the value per SNP of the original GWAS, values comparable to the results for the original 0.5 threshold (Appendix figures A.9B and A.10B). The NON NEURAL and NON ASSOCIATED partitions showed CoDs per SNP that were both ∼4.5 the value per SNP of the original GWAS.

Sensitivity analysis in other base-target sample pairs

Comparing the main results with those obtained in other large PGC cohorts at the original p=0.5 threshold, one finds that the original leave-one-out GWAS coefficient of determination (CoD) for schizophrenia in each sample varied between 7.67% (95% CI 6.55; 8.82) in the *celso*, and 11.08% (95% CI 10.36; 11.81) in the *clz2a* cohorts (Appendix figures A.5A, and A.1A), as compared to 9.85% in *xs234*. The CoD per SNP varied between 6.91 and 6.33 $\times 10^{-7}$ (Appendix figures A.5B, and A.1B), both close to the headline figure of 6.13 in *xs234*.

The NEURAL SIGNIFICANT partitions had CoDs of 2.63% (95% CI 1.95; 3.32) in the *celso*, and of 3.18% (95% CI 2.77; 3.58) in the *clz2a* cohorts, respectively (it was 3.72% in *xs234*). The CoD per SNP equated to 38.36 and 39.08 $\times 10^{-7}$, respectively, which are between 7 and 10 times the value per SNP of the original GWAS (Appendix figures A.5B, and A.1B), close to *xs234* ratios. NON NEURAL and NON ASSOCIATED partitions both had CoDs per SNP ∼30 in *celso*, and between 27 and 30 for *clz2a*, respectively, which are between 4 and 5 times the value per SNP of the respective original GWAS (Appendix figures A.7B, A.8B, A.3B, and A.4B).

3.4.1.3 Coefficients of determination for the original GWAS PRS vs multivariable models in schizophrenia

In this section, I compare the CoDs for various multivariable models incorporating the *enhancer-based* and *residual* partition PRSs as separate predictors. The comparator, once more, is the original leave-one-out GWAS. As it was not clear if the *enhancer-based* and *residual* partition PRSs – used as independent predictors of schizophrenia – might show interactions, or if their relationship with the outcome would be linear, they were modelled in three ways:

- As a simple *logit* additive model, in the form of: $SCZ \sim TS\_ENH\_PRS + residual\_PRS$, where SCZ is schizophrenia (binary), and TS_ENH_PRS and $residual\_PRS$ are the continuous PRSs for each partition.
- As a *logit* additive model plus interactions, in the form of: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS$.
- As a *logit* additive model plus interactions, plus quadratic terms: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS + TS\_ENH\_PRS^2 + residual\_PRS^2$.

The main analysis, once more, focussed on the *xs234* European population.

As per the previous section, the original leave-one-out GWAS coefficient of determination for schizophrenia in this sample, based on ∼161k SNPs, equated to 9.85% (95% CI 8.72; 11.01). The simple *logit* additive model explained respectively 10.26% (95% CI 9.11; 11.45) and 10.37% (95% CI 9.22; 11.56) of the adjusted variance utilising the NEURAL SIGNIFICANT, or the NEURAL SIGNIFICANT WITHIN GRBS partitions, respectively (Figure 3.4). This is a small increment, despite a reduction in SNPs accounted for by the two models, with ∼157K and ∼159K, respectively. The point values for the CoD for the two non-significant, control partitions were lower, at 9.81 and 9.90% for the NON NEURAL and the NON ASSOCIATED partitions, respectively (Figure 3.5).

For the *logit* additive model plus interactions, the CoD values were comparable to the simple additive model's, with point values gaining 0.01 to 0.05 percentage points in the

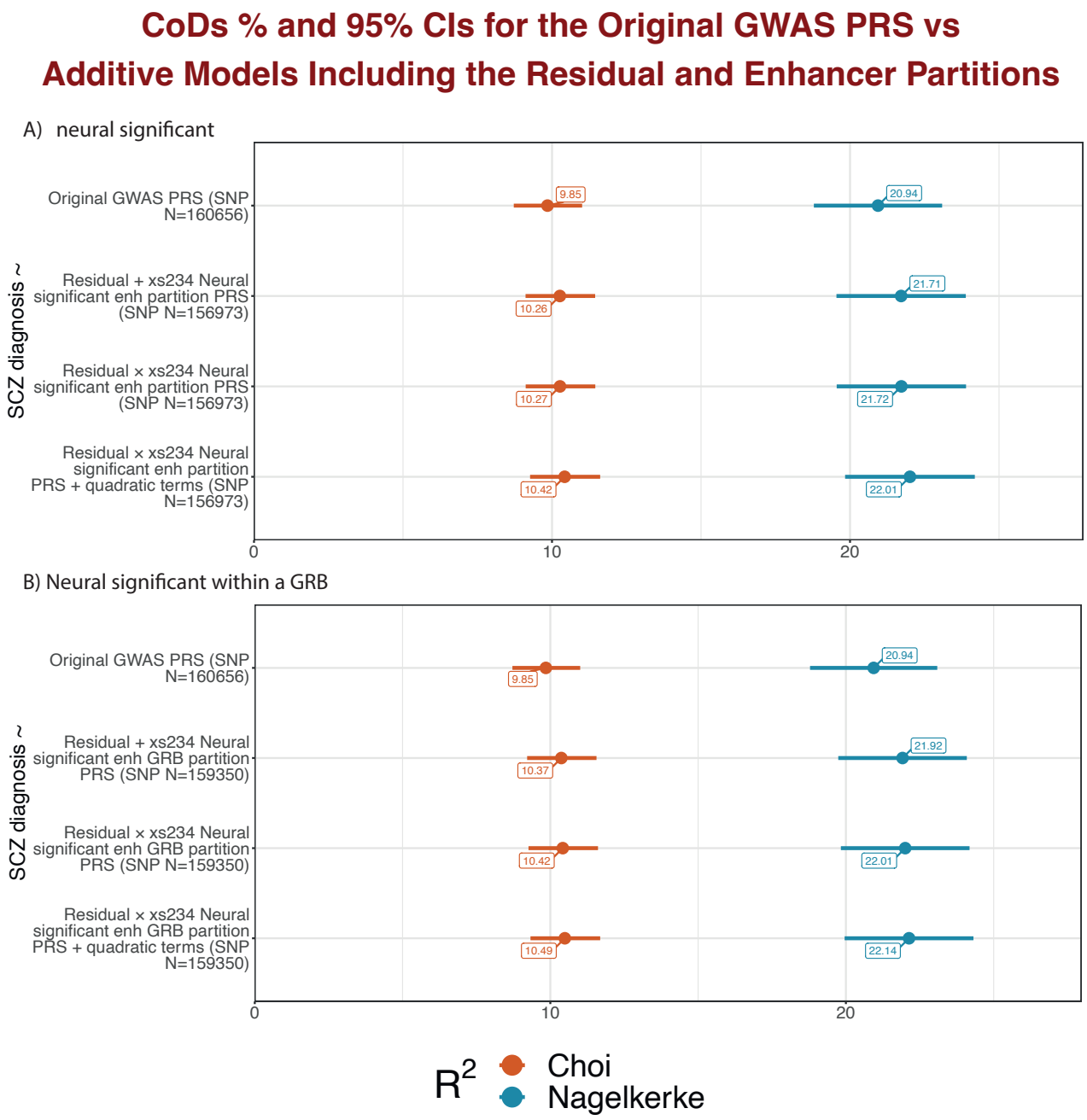


Figure 3.4: Coefficients of determination for schizophrenia for the original GWAS PRS vs multivariable models in the *xs234* cohort – significant partitions.
The figure describes the proportion of the variance of schizophrenia explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original LOO GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $SCZ \sim TS\_ENH\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS + TS\_ENH\_PRS^2 + residual\_PRS^2$.
Panels A) and B) represent the CoDs for each genomic partition for the NEURAL SIGNIFICANT and NEURAL SIGNIFICANT WITHIN GRBS lists, respectively.
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

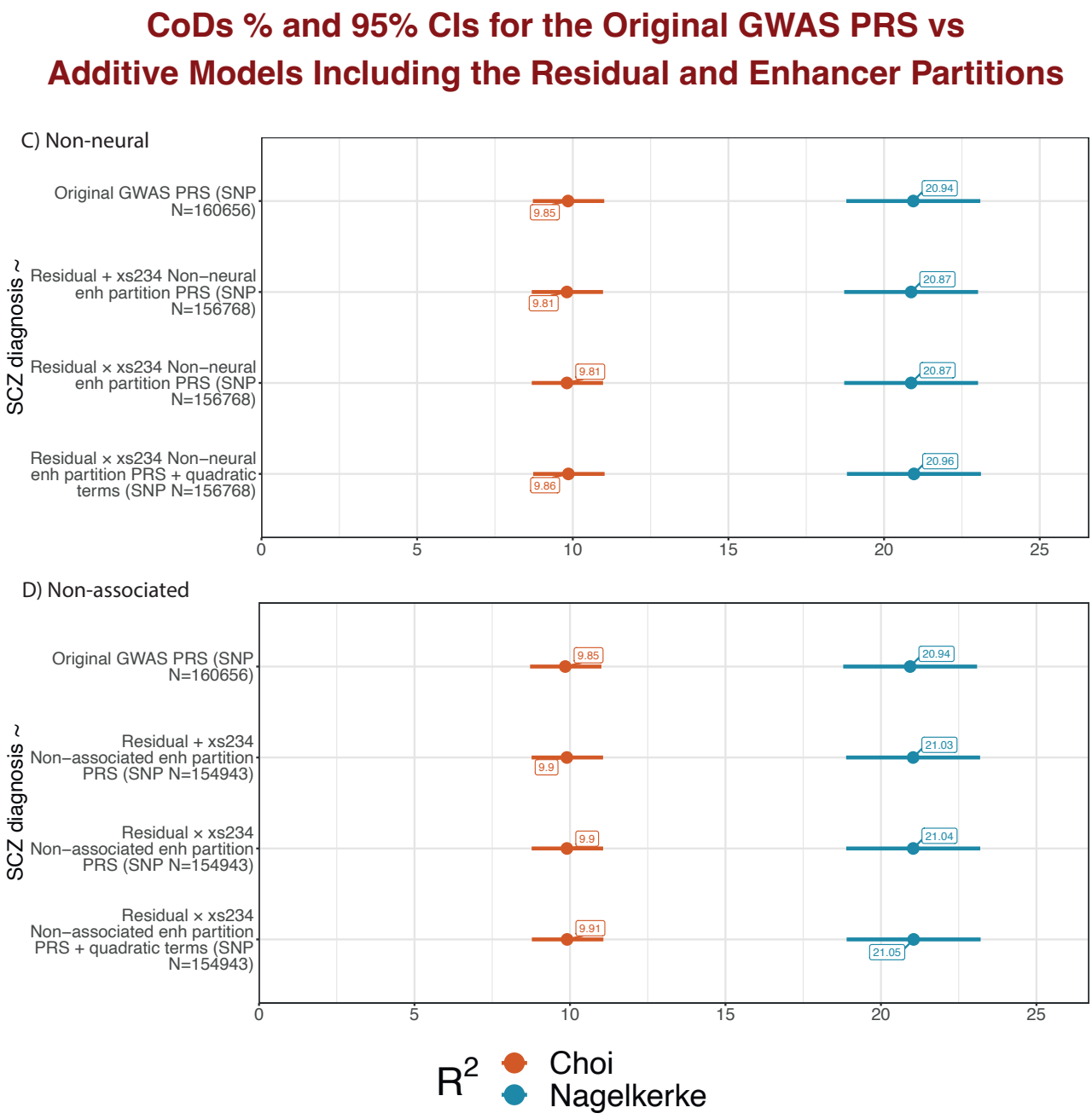


Figure 3.5: Coefficients of determination for schizophrenia for the original GWAS PRS vs multivariable models in the *xs234* cohort – non-significant partitions.
The figure describes the proportion of the variance of schizophrenia explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original LOO GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $SCZ \sim TS\_ENH\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $SCZ \sim TS\_ENH\_PRS \times residual\_PRS + TS\_ENH\_PRS^2 + residual\_PRS^2$.
Panel C) and D) represent the CoDs for each genomic partition for the NON NEURAL and NON ASSOCIATED lists, respectively.
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

NEURAL SIGNIFICANT, or the NEURAL SIGNIFICANT WITHIN GRBS partitions, respectively (Figure 3.4). Point values for the CoD for schizophrenia for the two non-significant, control partitions remained unchanged in the interaction models as compared to the simple additive model (Figure 3.5).

Finally, for the *logit* additive model plus interactions plus quadratic terms, the CoD values showed a very small further increase, reaching 10.42 and 10.49% in the NEURAL SIGNIFICANT, or the NEURAL SIGNIFICANT WITHIN GRBS partitions, respectively (Figure 3.4). Point values for the CoD for schizophrenia for the two non-significant, control partitions also modestly climbed in the interaction + quadratic terms models as compared to the simple additive model (Figure 3.5).

In conclusion, according to the point estimate, the 'best' tissue-specific-enhancer pPRS-based model for this analysis appears to be that based on NEURAL SIGNIFICANT WITHIN GRBS; this partition, based on an additive model plus interactions plus quadratic terms, shows a CoD gain of 6.5% as compared to the original leave-one-out GWAS coefficient of determination for schizophrenia.

Sensitivity analysis in *xs234* at the *p*-value threshold of 0.05

I then tested whether comparing PRSs between partitions at a different *p*-value threshold of 0.05 showed major differences from the main analysis. The figures for each enhancer list for this analysis are in Appendix Figures A.9 to A.12.

In comparison to the higher CoD for schizophrenia from the original LOO GWAS at this threshold of 10.28% (95% CI 9.13; 11.47), that of all three multivariable models showed a modest reduction, to point values ranging between 10.09 and 10.15% for the NEURAL SIGNIFICANT list, while the same coefficients for the NEURAL SIGNIFICANT WITHIN GRBS list showed a modest increase, to point values ranging from 10.44 to 10.45% (Figures A.9D and A.10D).

Non-significant lists both showed reductions in the CoD for schizophrenia, with point values ranging from 9.86 to 10.01% for both the NON NEURAL and the NON ASSOCI-

ATED list within GRB lists (Figures A.11D and A.12D).

Sensitivity analysis in other base-target sample pairs

Appendix Figures A.1 to A.8 in Panel D show the results of the same analysis, at the original threshold of 0.5, for the *celso* and *clz2a* additional PGC samples. For both additional cohorts, the two significant enhancer lists showed the same small gradual increases as the *xs234*. Taking the NEURAL SIGNIFICANT list, within *celso*, the CoD for the original LOO GWAS was 7.67%, that for the additive model climbed to 7.96%, and for the model with interactions and quadratic terms it climbed further to 8.00% (Figure A.5D). For NEURAL SIGNIFICANT ENHANCERS WITHIN GRBS within *celso*, the CoD for the additive model climbed to 8.17%, and for the model with interactions and quadratic terms it climbed further to 8.18% (Figure A.6D), **an increase of 6.6%**. Within *clz2a*, the CoD for the original LOO GWAS was 11.08%, that for the additive model climbed to 11.21%, and for the model with interactions and quadratic terms it climbed further to 11.23% (Figure A.1D). For NEURAL SIGNIFICANT ENHANCERS WITHIN GRBS within *clz2a*, the CoD for the additive model climbed to 11.17%, and for the model with interactions and quadratic terms it climbed further to 11.19% (Figure A.2D), **an increase of 1%**.

Both non-significant lists showed inconsistent results between the *celso* and the *clz2a* samples, with decreased (*clz2a*) and increased (*celso*) CoDs for the multivariable models as compared to the original LOO GWAS (Figures A.3D, A.4D, A.7D, A.8D).

3.4.1.4 Coefficients of determination for *enhancer-based* partitions in schizophrenia

In this section, I tested whether – by multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by either the *effect size* of the tissue-specific enhancer, or by its tissue-specific expression – the PRS calculated using these statistics explained more of the adjusted variance for schizophrenia. Therefore, I calculated three pPRSs for each enhancer-based partition, for each base/target pair:

- The PRS based on the original GWAS OR, for comparison;

- OR_{ES}: A PRS where the OR was multiplied by each enhancer's effect size (or ES, the measure of association of the enhancer with its top target gene). The adjustment was done using equation 3.1, in order to maintain a value on the same scale as the original OR.
- $OR_{TS_{exp}}$: A PRS where the OR had been adjusted for each hosting enhancer's neural-specific expression value in tpm. The adjustment was done using equation 3.2, in order to maintain a value on the same scale as the original OR.

Of note, enhancer-specific measures were only available for enhancers annotated within the AR+C, and A) with tissue-specific expression in neural tissue > 0 for neural-specific expression; B) with evidence of significant contact with at least one nearby promoter for effect size. For those enhancers without one of these measures available (e.g., NON NEURAL ENHANCERS all had no neural expression available), the original OR value was used. Therefore, the results for the negative lists for this analysis were not deemed relevant, and won't be reported. Once more, the main analysis focussed on the *xs234* European population.

Both significant lists showed no improvement in CoDs for schizophrenia calculated using the OR_{ES} or the $OR_{TS_{exp}}$, as compared to the original OR. The ES-enhanced list even saw a small drop in its CoD (Figure 3.6).

Sensitivity analysis in *xs234* at the *p*-value threshold of 0.05

I tested whether comparing PRSs between *enhancer-based* partitions at a different *p*-value threshold of 0.05 produced different results. Confirming the main results, both significant lists showed no change in CoDs for schizophrenia calculated using the OR_{ES} or the $OR_{TS_{exp}}$, as compared to the original OR, or a very small drop (Figures A.9C and A.10C).

Sensitivity analysis in other base-target sample pairs

Panel C of Appendix Figures A.1, A.2, A.5, and A.6 shows the results of the same analysis, at the original threshold of 0.5, for the *celso* and *clz2a* additional PGC samples. The model developed using the NEURAL SIGNIFICANT list in the *celso* cohort saw a small drop

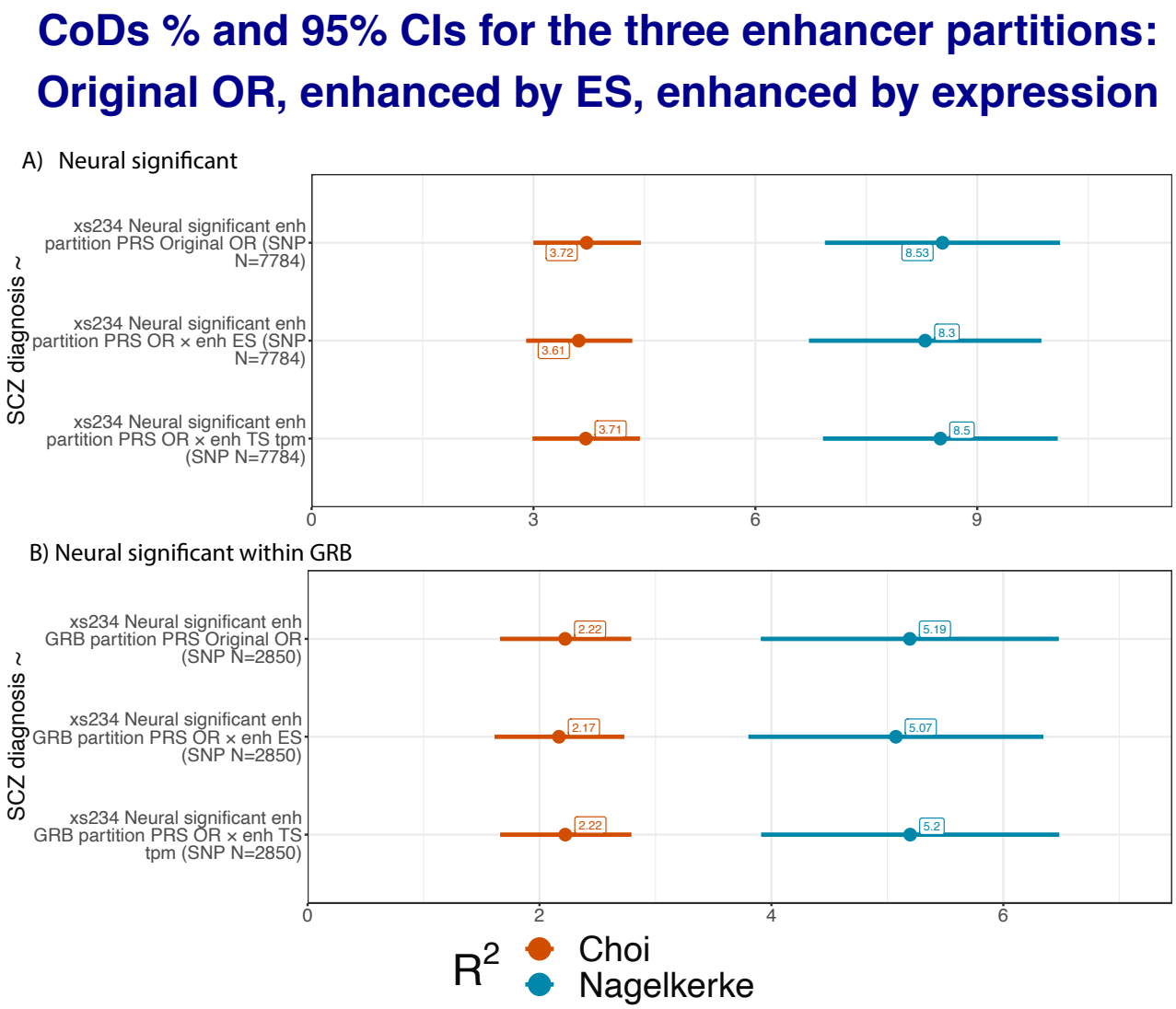


Figure 3.6: Coefficients of determination for schizophrenia for enhancer-based partitions in the *xs234* cohort – significant partitions.
The figure describes the proportion of the variance of schizophrenia explained by the genetic factor for three enhancer-based PRSs – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. **Panels A) and B)** show the CoDs for each genomic partition for the NEURAL SIGNIFICANT and for the NEURAL SIGNIFICANT WITHIN GRBS lists, respectively.
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

in the CoD for schizophrenia for both alternative OR measures, while the model developed with the NEURAL SIGNIFICANT WITHIN GRBS list saw a very small increase and a small decrease for the CoDs calculated using the OR_{ES} or the $OR_{TS_{exp}}$, as compared to the original OR, respectively (Panel C of Appendix Figures A.5 and A.6). The models developed using the *clz2a* cohort for both significant enhancer lists saw a small decrease and no change for the CoDs calculated using the OR_{ES} or the $OR_{TS_{exp}}$, as compared to the original OR, respectively (Panel C of Appendix Figures A.1 and A.2).

3.4.1.5 Can tissue-specific, *enhancer-based* partitioned PRSs help stratify people at risk for schizophrenia?

The final application of pPRSs for schizophrenia was to test them as adjuncts to canonical PRSs to stratify people for schizophrenia risk. To do so, I created 'double' quantile plots. Regular quantile plots are plots where the risk for a condition, in this case schizophrenia, is expressed as an OR on the *y* axis, while the discrete *x* axis represents population quantiles based on their PRS for the disease. Usually, and again in this case, the ORs for schizophrenia for higher quantiles are relative to the OR for schizophrenia of the first quantile, which is the reference. These plots are useful to compare ORs for the condition between those at lowest genetic risk (lowest PRS quantile), and those at the highest.

'Double' quantile plots work exactly as regular quantile plots, representing the OR for schizophrenia on the *y* axis, and population quantiles on the *x* axis. However, for each original PRS quantile, the population is further subdivided into three sub-quantiles, based on each participant's NEURAL SIGNIFICANT ENHANCERS partition pPRS. As shown in Figure 3.7, the plot represents both population quantiles – in brick colour – and sub-quantiles – in darkening shades of blue for sub-quantiles 1 to 3. It is easy to see in the figure that – while stratifying the *xs234* cohort by main PRS quantiles clearly separates people by their schizophrenia ORs, with people in the third quantile having a schizophrenia OR of > 7 as compared to those in the first – further stratifying each quantile by NEURAL SIGNIFICANT ENHANCER pPRS does not provide any additional benefit, as the three sub-quantiles do not

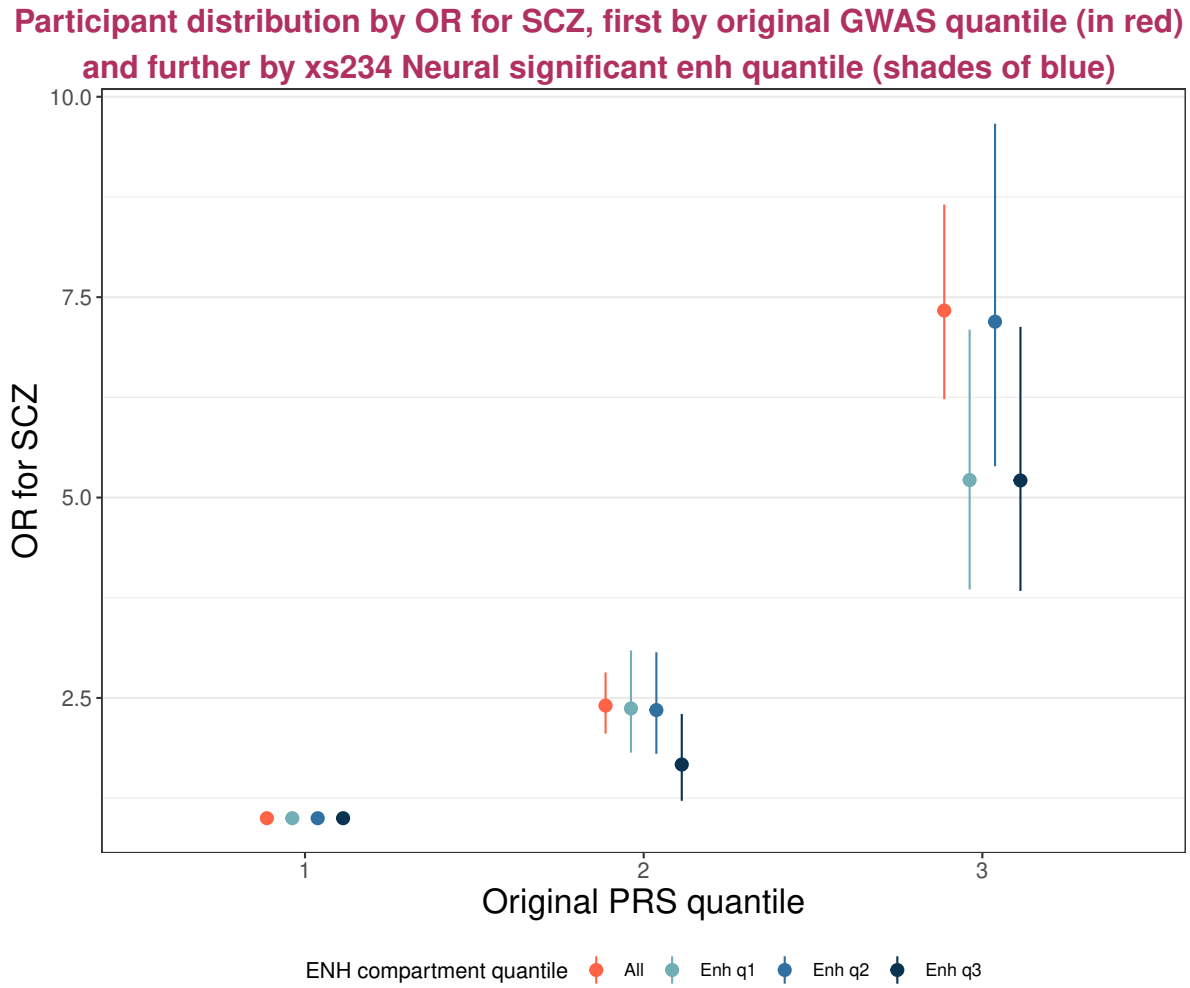


Figure 3.7: 'Double' quantile plot, expressing odds ratios (ORs) for schizophrenia in the *xs234* cohort.
The quantile plot shows, in brick red, the odds ratios and 95% confidence intervals for schizophrenia for three quantiles of original GWAS PRS, from 1 (reference, and lowest PRS) to 3. For each original GWAS PRS quantile, the population was then subdivided into three further quantiles, based on each participant's NEURAL SIGNIFICANT partition pPRS. The enhancer-based quantiles (from Enh q1 to Enh q3) are plotted as ORs and 95% confidence intervals in darkening shades of blue. .
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

appear to show increasing risk from lower to higher pPRS values.

3.4.2 Cardiac tissue-specific enhancers and variance explained in HCM

Analyses of HCM and cardiac tissue-specific enhancers were performed on the largest GWAS to date for the condition as the base (see section 3.2.2 and Tadros et al., 2023), and using UK Biobank genotypes as the target dataset (see section 3.2.3 and Bycroft et al., 2018). A sensitivity analysis including results obtained running the pipeline at the additional threshold of p=0.05 (instead of the main p=0.5) for same base/target populations can be found in the Appendix, in section A.3. A further sensitivity analysis was performed using the unpublished Royal Brompton leave-one-out GWAS from Tadros et al., 2023 as the base, and the Royal Brompton HCM cohort as the target dataset.

The analyses started from the following enhancer lists:

1. CARDIAC SIGNIFICANT ENHANCERS: $\sim 9K$ cardiac-expressed enhancers, with significant co-expression with at least one promoter, and with evidence of 3D contact between the enhancer and the promoter from the AR+C.
2. NON CARDIAC ENHANCERS, $\sim 41K$ enhancers with no cardiac expression – not necessarily in a significant enhancer-promoter pair (negative control 1).
3. NON-ASSOCIATED ENHANCERS: $\sim 34K$ FANTOM5 enhancers not associated to a gene – not necessarily with any cardiac expression (negative/neutral control 2).

3.4.2.1 Patient and SNP selection for the UK Biobank target cohort

The HCM base GWAS based on European populations contained information about 5,606,779 SNPs, based on 900 HCM cases and 68,3593 controls (Tadros et al., 2023). The target and base files were quality controlled, which entailed removing rare variants (MAF $< 1\%$), and those with INFO scores < 0.8 (low imputation quality), as well as filtering out SNPs in controls outside Hardy-Weinberg Equilibrium or with high missingness. Then, considering the specific base and target datasets, the two datasets were harmonised by strand flipping and removing mismatching SNPs. See the methods (section 3.3.1) for details.

After base GWAS QC, information remained about 5,600,542 SNPs, which further

reduced to 236,509 following clumping. The original UKBB cohort included $\sim 500K$ volunteers, of which 597 cases of HCM, before any filtering. Each participant had ∼96 million imputed SNPs available for testing. Following European ancestry selection, as well as QC, 413,415 participants remained, of which 455 cases of HCM. Only SNPs in the clumped base GWAS or within the selected enhancer lists were retained, resulting in 289,499 SNPs per participant.

In terms of tissue-specific lists, when splitting the base GWAS into an *enhancer-based* and *residual* partitions, the total sum of SNPs in these partitions did not equal to the QCed, clumped original GWAS partition, as the splitting process was doped by prioritising SNPs in enhancer partitions by temporarily reducing their *p*-value before joint clumping. As an example, here I present data for one such lists, the CARDIAC SIGNIFICANT ENHANCERS, as resulting from the specific base/target QC:

- Number of HCM GWAS SNPs in the *enhancer-based* partition before clumping: 8,901.
- Number in the corresponding *residual* partition before clumping: 5,591,641. The sum of the two before clumping still equalled 5,600,542 SNPs in the original GWAS partition before clumping.
- After clumping, there were 5,544 SNPs left in the CARDIAC SIGNIFICANT ENHANCERS partition, preserving 62% of pre-clumping SNPs.
- After clumping, there were 245,233 SNPs left in the corresponding *residual partition*, preserving 4.4% of pre-clumping SNPs.

Similar results were obtained for all enhancer lists. The exact numbers of SNPs in each partition are presented in each section below.

3.4.2.2 Coefficients of determination for the main genomic partitions in HCM

First, I examined how much of the variance for the HCM phenotype could be explained by the three main genomic partitions: the *original*, clumped GWAS, for comparison; the *enhancer-based* partitions; and the *residual* partition for each enhancer list. The results in Figure 3.8 show that the original GWAS CoD for HCM in this sample, based on $\sim 166K$

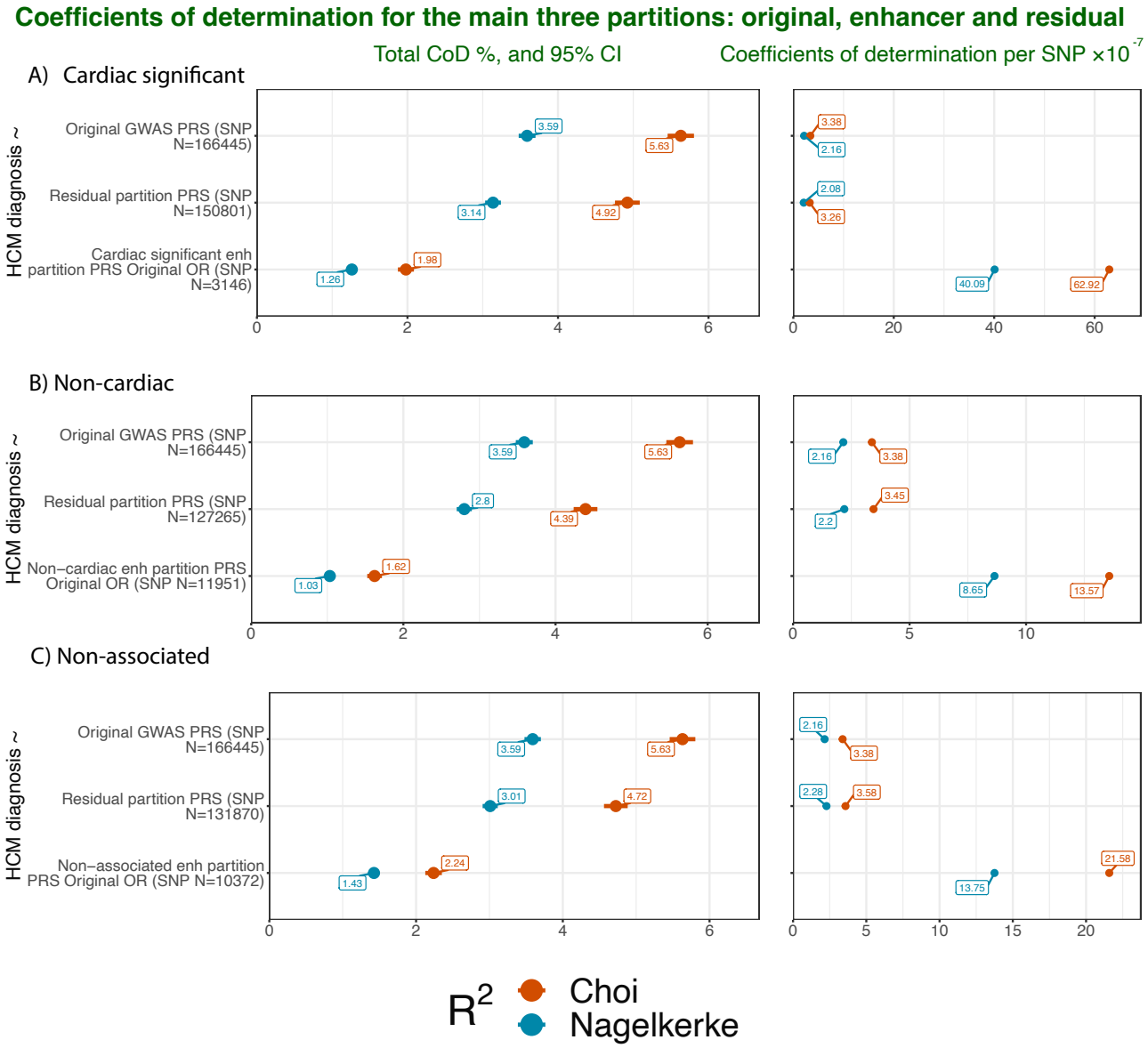


Figure 3.8: Coefficients of determination for HCM for the three main partitions, original GWAS, *residual* and *tissue-specific* enhancers, in the *UKBB* cohort.
The figure describes the proportion of the variance of HCM explained by the genetic factor for each PRS for the three main genomic partitions – original GWAS, *residual* and *tissue-specific* enhancers, in the *UKBB* cohort.
In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Each plot on the left shows the overall CoD % and 95% confidence interval, and on the right the corresponding point value adjusted per SNP ($\times 10^{-7}$). **Panels A) to C)** show the CoDs for each genomic partition for the CARDIAC SIGNIFICANT, NON CARDIAC, and for the NON ASSOCIATED lists, respectively.
Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

SNPs, equated to 5.63% (95% confidence interval (CI): 5.46; 5.81). The CoD per SNP equated to 3.38×10^{-7}. These values did not differ between enhancer lists. The CARDIAC SIGNIFICANT ENHANCERS partition had a CoD of 1.98% (95% CI 1.87; 2.08). The CoD per SNP equated to 62.92×10^{-7}, which is ~ 19 times the value per SNP of the original GWAS (Figure 3.8A). The control lists (NON NEURAL and NON ASSOCIATED ENHANCERS) showed lower explained variance per SNP as compared to the CARDIAC SIGNIFICANT ENHANCERS partitions, with CoDs per SNP of 13.57 and 21.58 $\times 10^{-7}$, respectively. These values are between 4 and 6 times the value per SNP of the original GWAS (Figures 3.8B) and 3.8C).

Sensitivity analysis on the HCM Royal Brompton cohort

The results in Figure 3.9 show that the original leave-one-out GWAS CoD for HCM in this sample, based on $\sim 63K$ SNPs, equated to 5.37% (95% confidence interval (CI): 4.09; 6.69). The CoD per SNP equated to 8.48×10^{-7}. The CARDIAC SIGNIFICANT ENHANCERS partition had a CoD of 1.09% (95% CI 0.50; 1.69). The CoD per SNP equated to 40.36×10^{-7}, which is ~ 5 times the value per SNP of the original GWAS (Figure 3.9A). The control lists (NON NEURAL and NON ASSOCIATED ENHANCERS) showed lower explained variance per SNP as compared to the CARDIAC SIGNIFICANT ENHANCERS partitions, with CoDs per SNP of 26.02 and 33.34 $\times 10^{-7}$, respectively. These values are between 3 and 4 times the value per SNP of the original GWAS (Figures 3.9B) and 3.9C).

These values are similar to the main UK Biobank analysis, the two main differences being that in this cohort the CoD per SNP is lower for the CARDIAC SIGNIFICANT ENHANCERS partition, and all confidence intervals are wider, due to the smaller sample size (fewer controls). One other noticeable difference is the distribution of the raw Nagelkerke's pseudo-R^2 values, which this time sat on the right, rather than on the left of the adjusted CoDs. The reason for this discrepancy sits in the fact that, as discussed in section 3.2.2, the HCM Royal Brompton cohort included 448 patients with HCM and 1219 matched healthy controls, with a case:control ratio of 1:2.7, versus the ratio of 1:907 in UK Biobank; as a reminder, the assumed population prevalence of HCM in the population is 1:500. This

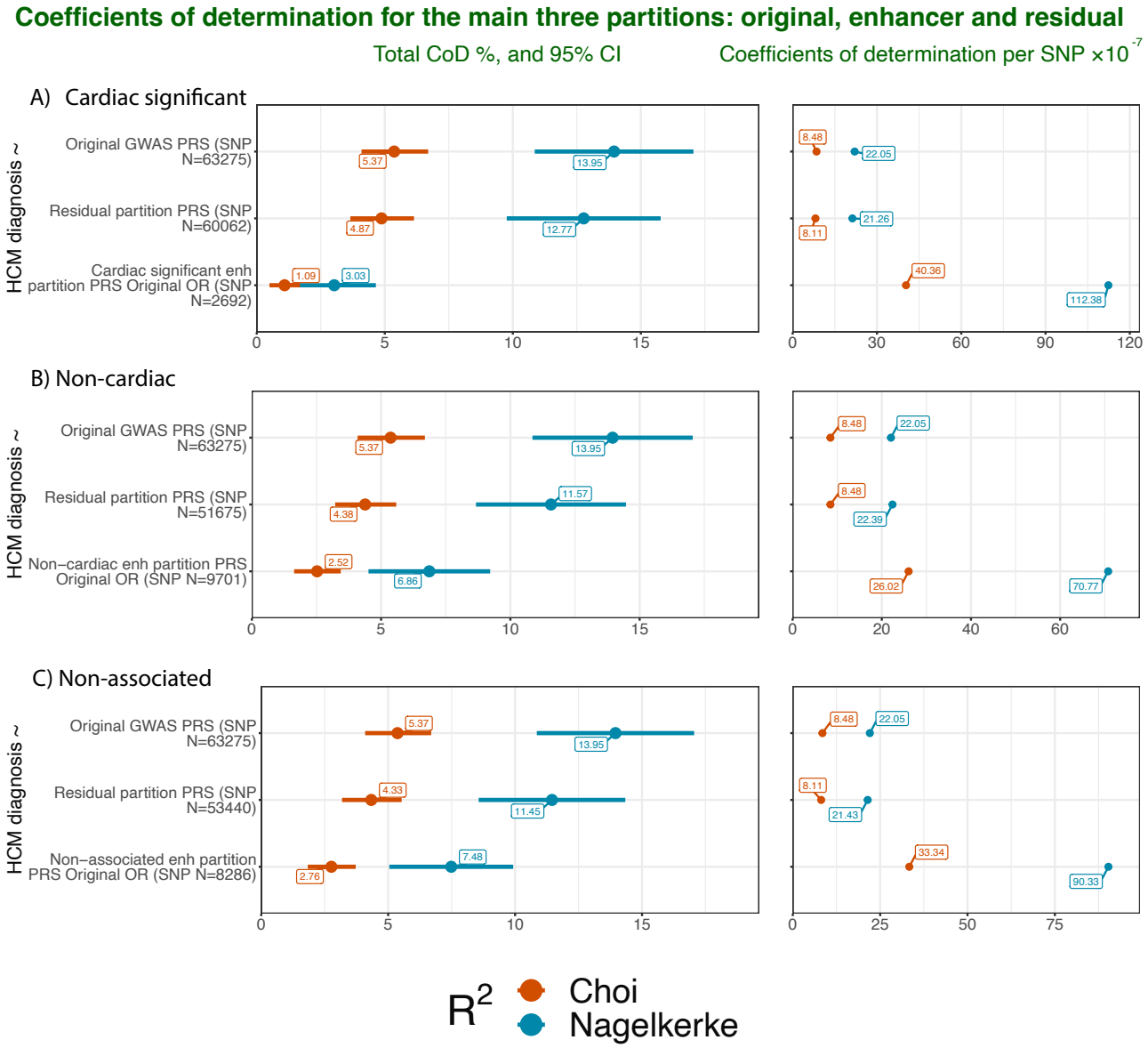


Figure 3.9: Coefficients of determination for HCM for the three main partitions, original GWAS, *residual* and *tissue-specific* enhancers, in the *Royal Brompton Hospital* HCM cohort.
The figure describes the proportion of the variance of HCM explained by the genetic factor for each PRS for the three main genomic partitions – original GWAS, *residual* and *tissue-specific* enhancers, in the *Royal Brompton Hospital* cohort.
In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Each plot on the left shows the overall CoD % and 95% confidence interval, and on the right the corresponding point value adjusted per SNP ($\times 10^{-7}$).
Panels A) to C) show the CoDs for each genomic partition for the CARDIAC SIGNIFICANT, NON CARDIAC, and for the NON ASSOCIATED lists, respectively.
Base data: RBH LOO HCM GWAS by Tadros et al., 2023. Target data: Royal Brompton Hospital cohort.

means that, while the UK Biobank was under-sampling the true prevalence, this RBH clinical sample was obviously enriched in patients, and therefore the raw model fit measures shifted to opposing directions. It was reassuring to find that the adjusted estimates were similar.

Sensitivity analysis at the *p*-value threshold of 0.05

Next, I compared PRSs between partitions while using a different threshold of 0.05 instead of 0.5. The figures for each enhancer list for this analysis are Appendix Figures A.13 to A.15. The smaller number of SNPs included at this lower threshold, $\sim 36K$, meant that the original GWAS coefficient of determination (CoD) for HCM in this sample equated to 4.72% (95% CI 4.56; 4.88), slightly lower than that at the original threshold of 0.5. The CoD per SNP equated to 13.08×10^{-7} (Appendix figure A.13A). At this threshold, the CARDIAC SIGNIFICANT ENHANCERS (502 SNPs) partition had a CoDs of 1.78% (95% CI 1.68; 1.88 – see Appendix figure A.13A). This partition showed a CoD per SNP that was ~ 27 times the value per SNP of the original GWAS, even higher than at the 0.5 threshold (Appendix figure A.13B). The NON NEURAL and NON ASSOCIATED partitions showed CoDs per SNP that were between 6 and 9 times the value per SNP of the original GWAS.

3.4.2.3 Coefficients of determination for the original GWAS PRS vs multivariable models in HCM

In this section, I compare CoDs for various multivariable models incorporating the *enhancer-based* and *residual* partition PRSs as separate predictors. As per the previous section, the coefficient of determination for HCM for the original GWAS in this sample, based on $\sim 166K$ SNPs, equated to 5.63% (95% confidence interval (CI): 5.46; 5.81). A simple *logit* additive model ($HCM \sim TS\_ENH\_PRS + residual\_PRS$) explained 6.32% (95% CI 6.13; 6.50) of the adjusted variance utilising the CARDIAC SIGNIFICANT ENHANCERS PRS (Figure 3.10). This is an increment as compared to the original GWAS PRS CoD.

Interestingly, the additive model based on the NON CARDIAC ENHANCER PRS

showed a significant **decrease** in CoD as compared to that based on the original GWAS PRS (5.26%; 95% CI 5.09; 5.43). The CoD for the additive model based on the the NON ASSOCIATED ENHANCERS partition was 6.08%, not dissimilar from that based on the original PRS (Figure 3.10).

Finally, I calculated additive models plus interactions, and additive models plus interactions, plus quadratic terms. These were calculated for all three enhancer-based compartments, CARDIAC SIGNIFICANT, NON CARDIAC and NON ASSOCIATED ENHANCERS. As shown in Figure 3.10, the CoDs for these interactive models showed an **increasing** trend, as compared to the original GWAS PRS CoD, reaching a value of 9.31% (95% CI 9.09; 9.52) for the model based on CARDIAC SIGNIFICANT ENHANCERS. Unexpectedly, the NON CARDIAC and NON ASSOCIATED ENHANCERS-based models showed similar increasing trends.

Sensitivity analysis on the HCM Royal Brompton cohort

As described earlier, and as shown in Figure 3.11, the coefficient of determination for HCM for the original GWAS in this sample equated to 5.37%. A simple *logit* additive model ($HCM \sim TS\_ENH\_PRS + residual\_PRS$) explained 5.30% (95% CI 4.03; 6.62) of the adjusted variance utilising the CARDIAC SIGNIFICANT ENHANCERS PRS, while it explained 5.43% and 5.65% of the adjusted variance for the NON CARDIAC ENHANCERS and for the NON ASSOCIATED ENHANCERS partitions, respectively.

Finally, additive models plus interactions, and additive models plus interactions, plus quadratic terms were calculated for all three enhancer-based compartments, CARDIAC SIGNIFICANT, NON CARDIAC and NON ASSOCIATED ENHANCERS. As shown in Figure 3.11, the CoDs for these interactive models showed very marginally **increasing** trends, as compared to the original GWAS PRS CoD, and even more so for the non-significant partitions.

Sensitivity analysis at the *p*-value threshold of 0.05

Appendix Figures A.13 to A.15 show the results for the same analyses as those shown in the previous paragraph, this time at a PRS threshold of 0.05 as a sensitivity ana-

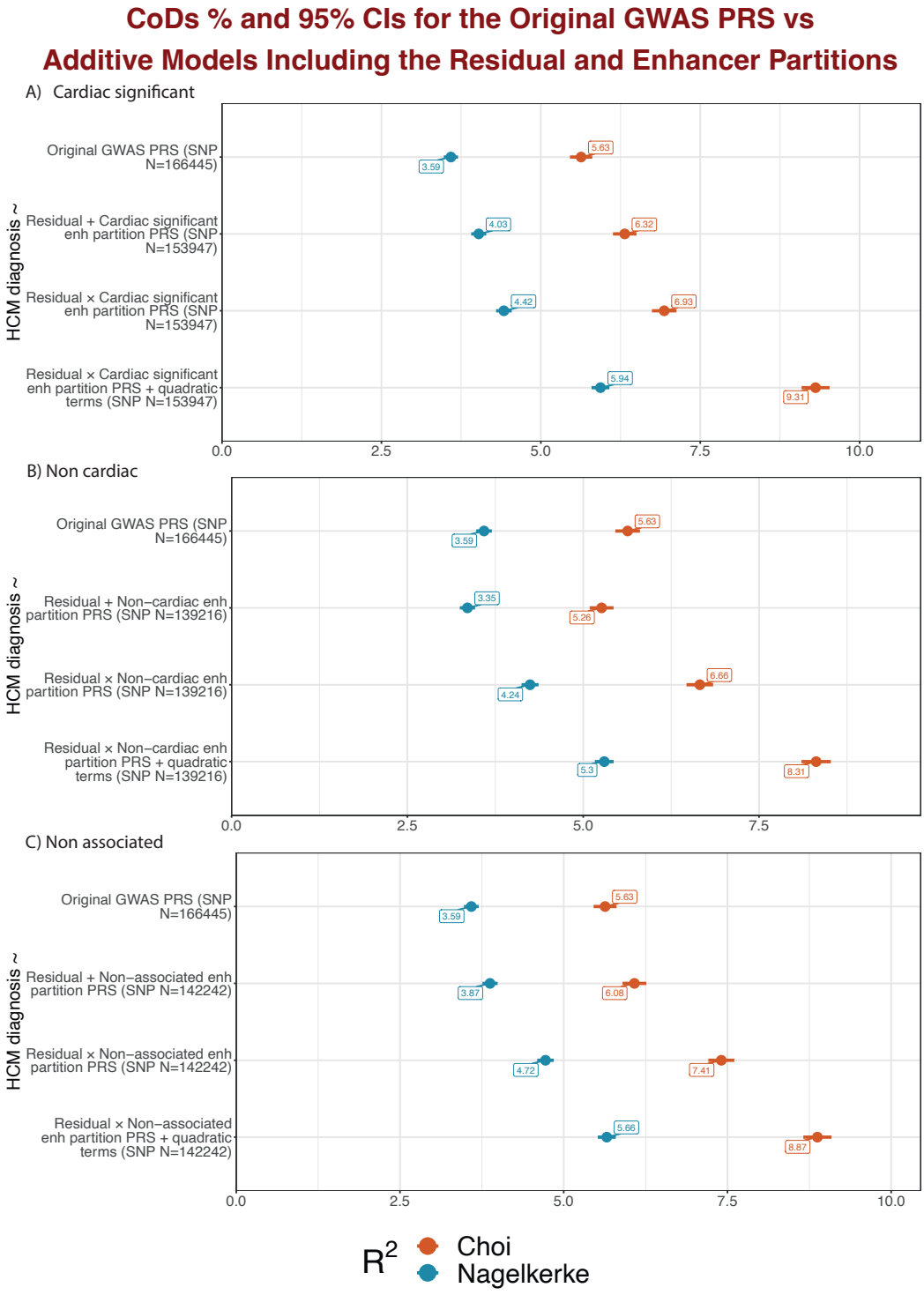


Figure 3.10: Coefficients of determination for HCM for for the original GWAS PRS vs multivariable models.

The figure describes the proportion of the variance of HCM explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $HCM \sim TS\_ENH\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $HCM \sim TS\_ENH\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $HCM \sim TS\_ENH\_PRS \times residual\_PRS + residual\_PRS^2 + TS\_ENH\_PRS^2$.

Panels A), B), and C) represent the coefficients of determination for each genomic partition for the CARDIAC SIGNIFICANT, NON CARDIAC, and NON ASSOCIATED lists, respectively.

Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

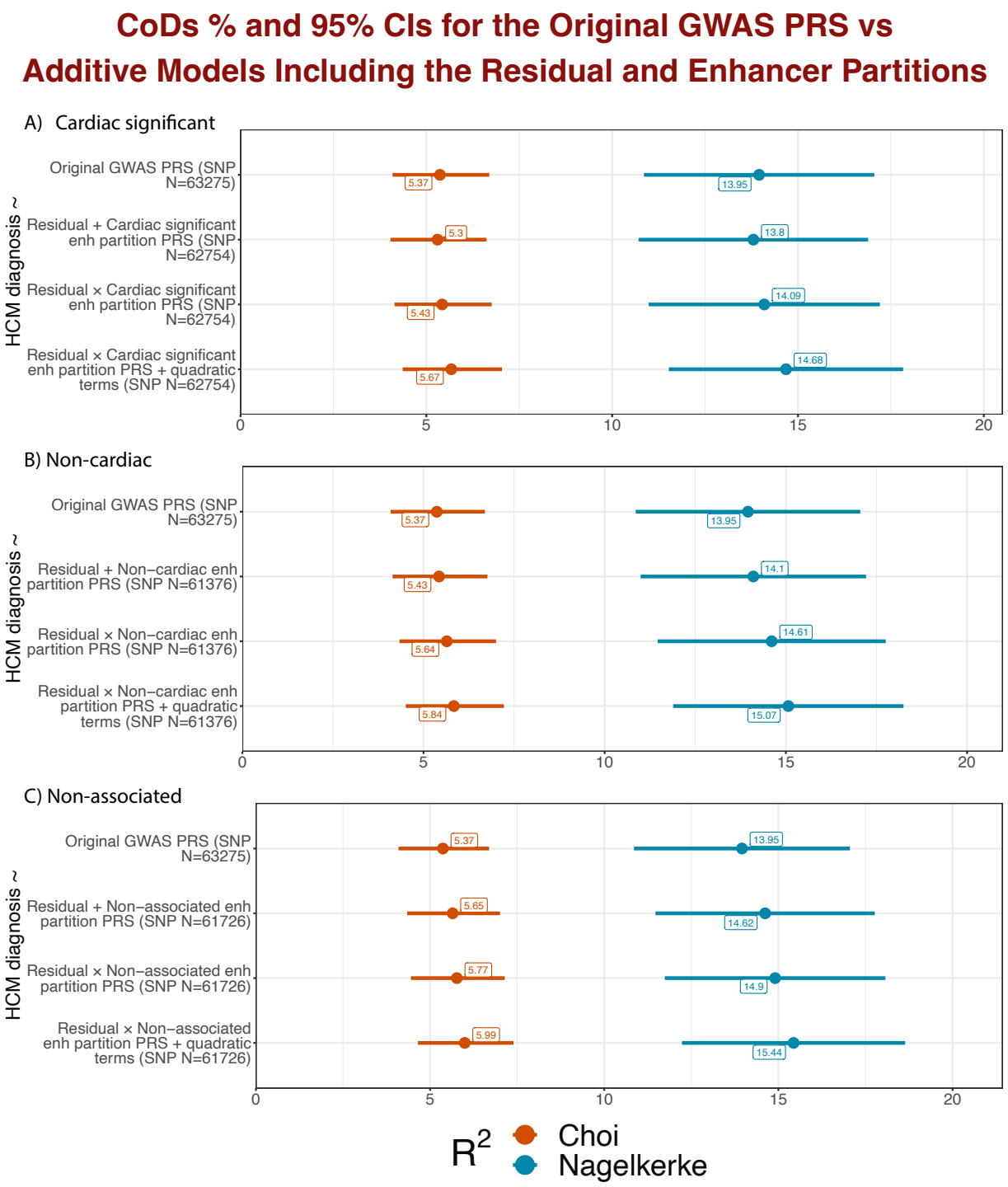


Figure 3.11: Coefficients of determination for HCM for for the original GWAS PRS vs multivariable models, in the *Royal Brompton Hospital* HCM cohort.
The figure describes the proportion of the variance of HCM explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original leave-one-out GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $HCM \sim TS\_ENH\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $HCM \sim TS\_ENH\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $HCM \sim TS\_ENH\_PRS \times residual\_PRS + residual\_PRS^2 + TS\_ENH\_PRS^2$.
Panels A), B), and C) represent the coefficients of determination for each genomic partition for the CARDIAC SIGNIFICANT, NON CARDIAC, and NON ASSOCIATED lists, respectively.
Base data: RBH LOO HCM GWAS by Tadros et al., 2023. Target data: Royal Brompton Hospital cohort.

lysis. The headline results from the previous paragraph are confirmed at this more stringent threshold: the additive, additive models plus interactions, and additive models plus interactions, plus quadratic terms, showed an increasing CoD pattern for all enhancer lists, with the exception of the NON CARDIAC ENHANCER PRS. This, as in the main analysis, showed a significant **decrease** in CoD as compared to that based on the original GWAS PRS for the simple additive model, which was then reversed in the additive model plus interactions, and additive model plus interactions, plus quadratic terms.

3.4.2.4 Coefficients of determination for *enhancer-based* partitions in HCM

In this section, I tested whether – by multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by either the *effect size* of the tissue-specific enhancer, or by its tissue-specific expression – the PRS calculated using these statistics explained more of the adjusted variance for HCM. Calculating pPRSs for CARDIAC SIGNIFICANT ENHANCERS based on three different OR measures (the original GWAS OR, the OR_{ES} or the $OR_{TS_{exp}}$ – see the methods section 3.3.1.2), the CoDs for HCM did not show any significant improvements (Figure 3.12).

Sensitivity analysis on the HCM Royal Brompton cohort

As shown in Figure 3.13, there was no significant difference when calculating pPRSs for CARDIAC SIGNIFICANT ENHANCERS based on the three different OR measures (the original GWAS OR, the OR_{ES} or the $OR_{TS_{exp}}$).

Sensitivity analysis at the *p*-value threshold of 0.05

Confirming the main results – comparing PRSs between *enhancer-based* partitions at a different *p*-value threshold of 0.05 – the CARDIAC SIGNIFICANT ENHANCER partition showed no change in CoD for HCM calculated using the OR_{ES} or the $OR_{TS_{exp}}$, as compared to the original OR, or a very small drop (Figure A.13C).

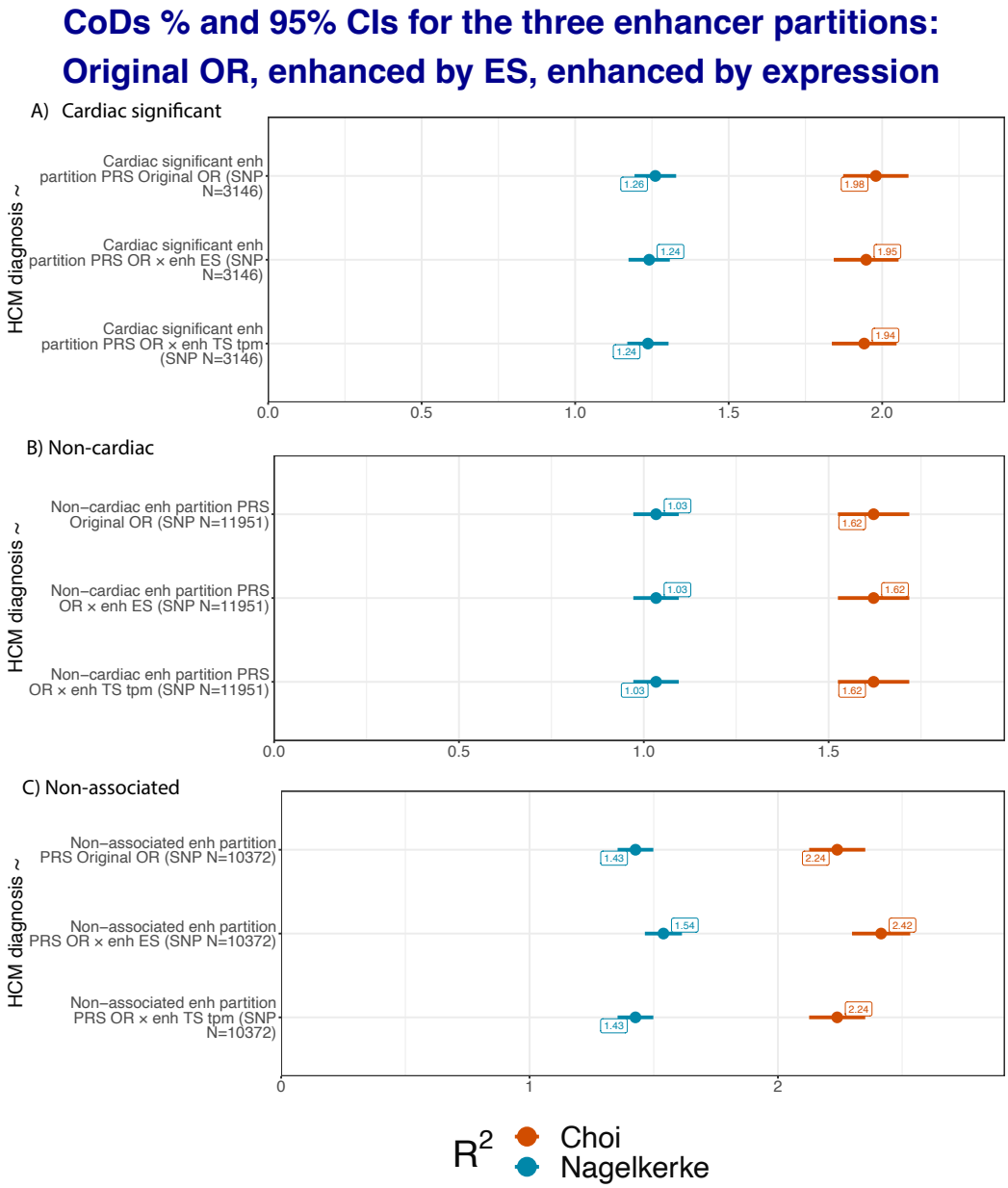


Figure 3.12: Coefficients of determination for HCM for enhancer-based partitions in the *UKBB* cohort.

The figure describes the proportion of the variance of HCM explained by the genetic factor for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Plots represent the CoD % and 95% confidence interval for the measure. **Panel A)** shows the CoD for each genomic partition for the CARDIAC SIGNIFICANT list. **Panel B)** shows the CoD for each genomic partition for the NON CARDIAC list. **Panel C)** shows the CoD for each genomic partition for the NON ASSOCIATED list.

Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

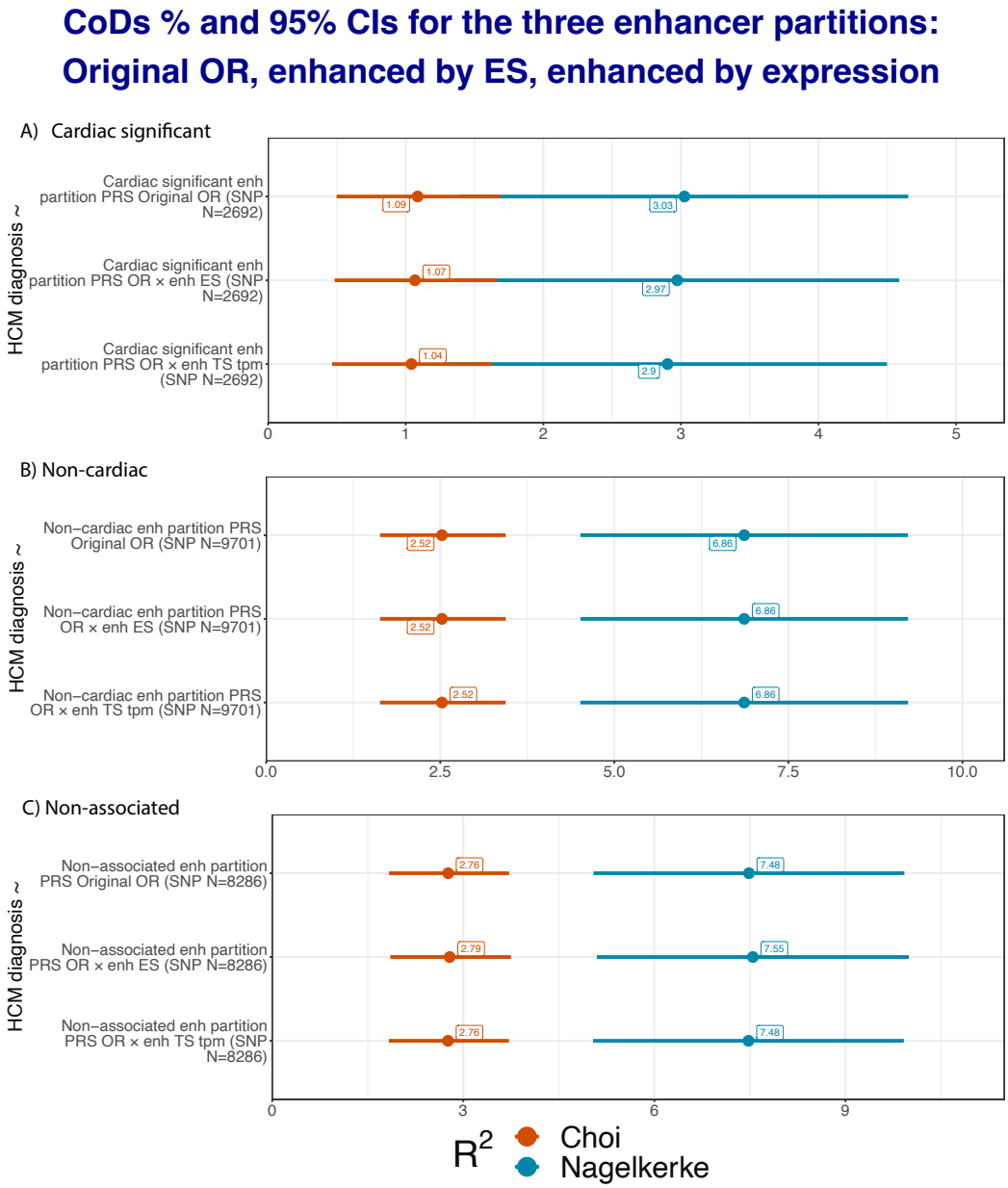


Figure 3.13: Coefficients of determination for HCM for enhancer-based partitions, in the *Royal Brompton Hospital* HCM cohort.
The figure describes the proportion of the variance of HCM explained by the genetic factor for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer.
In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Plots represent the CoD % and 95% confidence interval for the measure. **Panel A)** shows the CoD for each genomic partition for the CARDIAC SIGNIFICANT list. **Panel B)** shows the CoD for each genomic partition for the NON CARDIAC list. **Panel C)** shows the CoD for each genomic partition for the NON ASSOCIATED list.
Base data: RBH LOO HCM GWAS by Tadros et al., 2023. Target data: Royal Brompton Hospital cohort.

3.4.2.5 Can tissue-specific, *enhancer-based* partitioned PRSs help stratify people at risk for HCM?

The final application of pPRSs for HCM was to test them as adjuncts to canonical PRSs to stratify people for HCM risk. To do so, I plotted 'double quantile' plots for HCM. As a reminder, 'double quantile' plots work exactly as regular quantile plots, representing the OR for HCM on the *y* axis, and population HCM PRS quantiles on the *x* axis. However, for each original PRS quantile, the population is further subdivided into three sub-quantiles, based on each participant's CARDIAC SIGNIFICANT ENHANCERS partition pPRS. As shown in Figure 3.14, the plot represents both population quantiles – in brick colour – and sub-quantiles – in darkening shades of blue for sub-quantiles 1 to 3. It is easy to see in the figure that – while stratifying the *UKBB* cohort by main PRS quantiles clearly separates people by their HCM ORs, with people in the third quantile having a HCM OR of > 2.5 as compared to those in the first – further stratifying each quantile by CARDIAC SIGNIFICANT ENHANCER pPRS does not provide any additional benefit, as the three sub-quantiles do not appear to show increasing risk from lower to higher pPRS values.

3.5 Summary of findings

In this chapter I have developed 'partitioned' polygenic risk scores, or PRSs where two (or more) genomic compartments (e.g., the *tissue-specific enhancers* and the *residual* compartments) are considered separately for polygenic risk scoring – with special consideration for *enhancer-based* SNPs, which are prioritised – and where each model only includes LD-independent SNPs. Using logistic modelling and incorporating formulas to adjust the results for disease liability and for ascertainment, I have then calculated the amount of adjusted heritability explained by the original GWAS for schizophrenia and for HCM (h^2_{PRS}), as well as comparing this figure to the partitioned PRSs (h^2_{pPRS}), where the *enhancer-based* and the *residual* partition PRSs are used as separate predictors in a logistic model. I have then tested if multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by

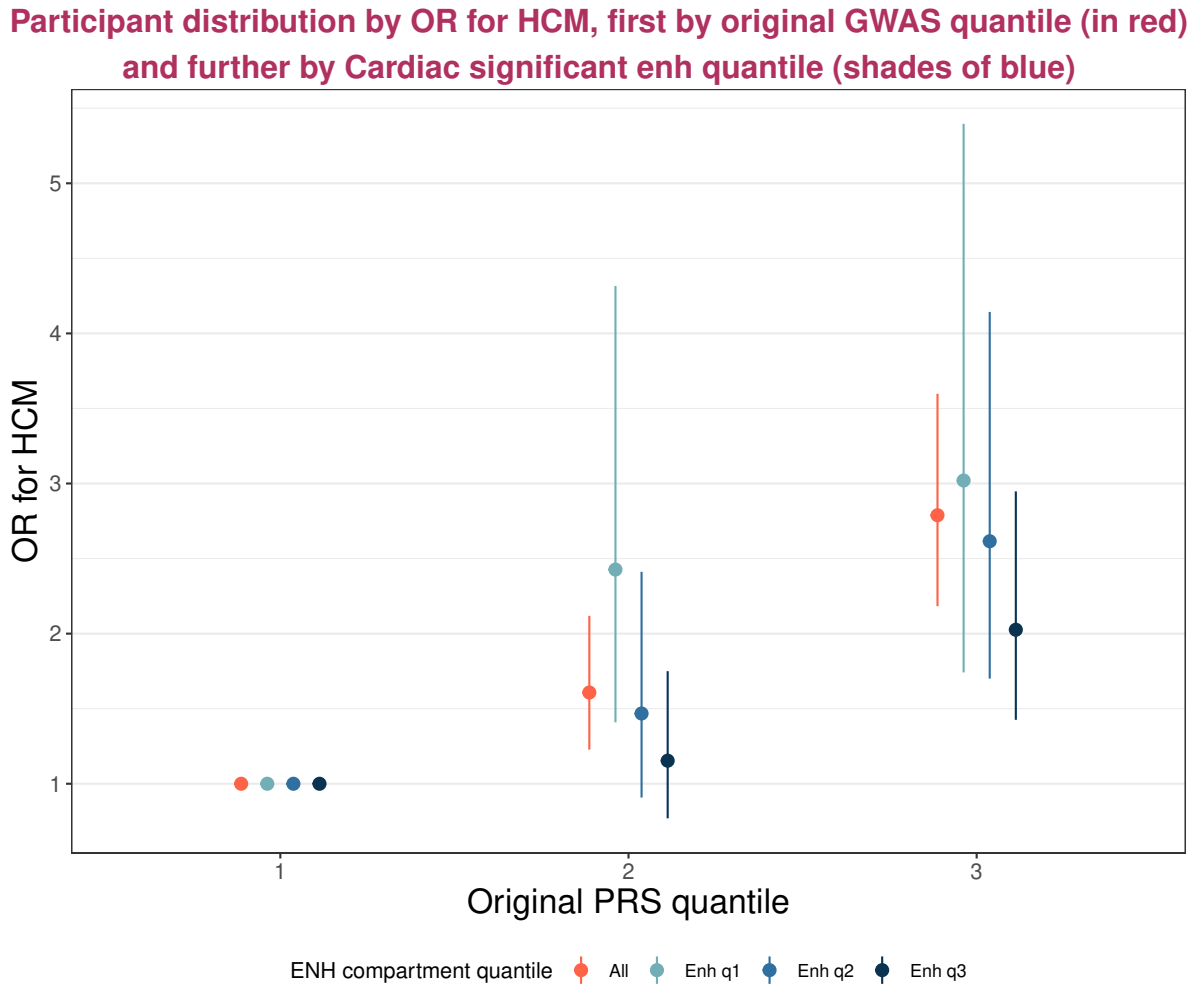


Figure 3.14: 'Double' quantile plot, expressing odds ratios (ORs) for HCM in the *UKBB* cohort. The quantile plot shows, in brick red, the odds ratios and 95% confidence intervals for HCM for three quantiles of original GWAS PRS, from 1 (reference, and lowest PRS) to 3. For each original GWAS PRS quantile, the population was then subdivided into three further quantiles, based on each participant's CARDIAC SIGNIFICANT partition pPRS. The enhancer-based quantiles (from Enh q1 to Enh q3) are plotted as ORs and 95% confidence intervals in darkening shades of blue.
Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

either an *effect size* of the tissue-specific enhancer, or by its neural expression, improved the overall disease h^2_{pPRS} as explained by partitioned PRS (pPRS).

The results for schizophrenia – based on the *xs234* cohort – show that:

1. The original leave-one-out-GWAS-based PRS – based on $\sim 161K$ SNPs – had a coefficient of determination (CoD) of 9.85% (equivalent to the schizophrenia h^2_{PRS} in this sample), while the NEURAL SIGNIFICANT ENHANCERS partition – based on $\sim 8K$ SNPs only – had a CoD of 3.72%; the NEURAL SIGNIFICANT WITHIN GRBS partition showed a CoD of 2.22% based on just $\sim 3K$ SNPs. The CoDs per SNP for enhancer-based partitions equated to between ∼8 and ∼13 times the value per SNP of the original GWAS (Figures 3.3A and 3.3B).
2. The use of multivariable logistic models to calculate the schizophrenia h^2_{pPRS} produced modest increases compared to the h^2_{PRS} figure. The best multivariable model, for the *logit* additive model plus interactions plus quadratic terms including NEURAL SIGNIFICANT WITHIN GRBS, reached a CoD of 10.49%, which, compared to the original leave-one-out GWAS coefficient of 9.85%, represents a 6.5% improvement. However, the quite wide confidence intervals do not allow to refute the null hypothesis (Figures 3.4A and 3.4B).
3. Using tissue-specific-variable-derived ORs for computing PRSs did not show any advantage over using the original OR measures for schizophrenia (Figures 3.6A and 3.6B).
4. Tissue-specific, *enhancer-based* partition PRSs do not appear useful, in addition to original PRS quantiles, to stratify populations for schizophrenia risk (Figure 3.7).

The results for HCM – based on both the *UK Biobank* and the *Royal Brompton Hospital* cohorts – show that:

1. In UKBB, the original GWAS-based PRS – based on $\sim 166K$ SNPs – had a coefficient of determination (CoD) of 5.63% (equivalent to the HCM h^2_{PRS} in this sample), while the CARDIAC SIGNIFICANT ENHANCERS partition – based on just $\sim 3K$ SNPs – had a CoD

of 1.26%. This equated to a CoD per SNP ~ 19 times the value per SNP of the original GWAS (Figure 3.8A). The RBH cohort showed similar results for the CoD of the LOO GWAS, however the CoD per SNP for CARDIAC SIGNIFICANT ENHANCERS equated to just ~ 5 times the value per SNP of the original leave-one-out GWAS (Figure 3.9A).

2. In UKBB, the use of additive logistic models to calculate the HCM h^2_{pPRS} showed a pattern different from schizophrenia, as demonstrated in Figure 3.10. For CARDIAC SIGNIFICANT ENHANCERS, as well as for NON ASSOCIATED partitions, CoDs for the simple additive, fully interactive, and fully interactive with quadratic terms kept improving in the amount of variance explained at each step.

   For NON CARDIAC ENHANCERS, a simple additive model showed a significant drop in h^2_{pPRS}, as compared to the HCM h^2_{PRS} in this sample, from 5.63 to 5.26%. The CoD then kept climbing in line with the other two for the fully interactive, and fully interactive with quadratic terms models.

3. In the RBH cohort, as demonstrated in Figure 3.11, CoDs for the simple additive, fully interactive, and fully interactive with quadratic terms also kept increasing in the amount of variance explained at each step, however the increases were much smaller. For CARDIAC SIGNIFICANT ENHANCERS, the percentage increase between the original and the fully interactive model including quadratic terms ($+5.6\%$) was similar to the gain seen for schizophrenia ($+6.5\%$). However, a significant difference with schizophrenia is that similar increases – and even larger ones – were seen for the non significant enhancer lists too.

4. Using tissue-specific-variable-derived ORs for computing PRSs did not show any advantage over using the original OR measures for HCM in either cohort (Figure 3.12A and 3.13A).

5. Tissue-specific, *enhancer-based* partition PRSs do not appear useful, in addition to original PRS quantiles, to stratify populations for HCM risk (Figure 3.14).

The findings are discussed in Chapter 5.

Chapter 4

Leveraging nonadditive disease inheritance models

4.1 Introduction

As discussed in the main introduction, GWASes are the current gold-standard technique for finding common SNP associations with *complex* disorders. Further, as we have seen in section 1.4.2, the replicability of GWAS findings relies on large population samples. For example, GWASes for a highly polygenic disorder such as schizophrenia did not have enough power to find strong, replicable results, up to the point when large consortia were founded, to allow for very large case-control samples to be formed (Psychiatric GWAS Consortium Coordinating Committee, 2009). Even today, with sample sizes including over $67K$ people with schizophrenia and over $90K$ controls (Trubetskoy et al., 2022), only a small

portion of schizophrenia's h^2 can be explained (see section 1.5.2). Further, when running GWAS, one can make assumptions on the most likely penetrance function, which implies a specific inheritance model, as we have seen in section 1.4.2.1. Choosing the correct inheritance model can be complicated: while with a Mendelian disorder one can look at a pedigree and study how the disorder spreads across families, and infer the best inheritance model, with *complex* disorders – as discussed in section 1.4.1.1 – there are several thousands to millions of SNPs tested against the same disease, all likely contributing to a very small extent to disease heritability. Therefore, the selection of an inheritance model – which can be done at the individual SNP level – can be a delicate endeavour.

One avenue that has not been often explored, is that of using different models of inheritance for separate genomic partitions within the same GWAS. As discussed in section 1.4.2.1, most existing GWASes use a *genetic-model-free* – see equation 1.3 – or an *additive* genetic model – see equation 1.4 – for all SNPs, thus assuming that genetic risk increases with each additional copy of an alternative allele for each and every SNP. This has been shown to be a reasonable assumption in most cases, and especially for highly polygenic conditions, as each SNP is supposed to only add a small amount of risk; further, this is what has worked in practice in most published GWASes (Bagos, 2013; Balding, 2006; Psychiatric GWAS Consortium Coordinating Committee, 2009). On the other hand, *dominance effects* are central to the study of model disease fitness by population geneticists, and across organisms and conditions the average dominance of mutation of small effects should be approximately one-quarter (Manna et al., 2011). Further, I have shown in Chapters 2 and 3 of this work that a method is available, that is effective at selecting genomic regions at higher-than-average (sometimes several times over) associated per-SNP schizophrenia heritability, i.e., neural-specific enhancers.

As a consequence, I have hypothesised that higher-priority variants (e.g., those falling inside neural-specific enhancers in the case of schizophrenia) might follow other models of inheritance, due to their higher-than-average disease heritability, and therefore a higher likelihood of being more disruptive. In this chapter, I will explore both *dominant*

and *recessive* inheritance models, as applied to tissue-specific enhancer-based genomic partitions, and their effects on schizophrenia GWAS results as compared to the canonical *additive* model. To avoid overlap between the GWAS development and PRS validation cohorts, I will calculate enhancer-based SNP association measures for schizophrenia (which I have called EP-WAS) on UK Biobank. I will then internally validate these measures of association on UK Biobank, and then externally validate them in the *xs234* PGC cohort.

4.2 Methods

4.2.1 Dominant and Recessive Schizophrenia EP-WAS in UK Biobank

To study the effects of both *dominant* and *recessive* inheritance models in schizophrenia GWAS, I developed an EP-WAS as follows:

1. UK Biobank genotypes (see section 3.2.3 and Bycroft et al., 2018) were first subjected to standard QC, including filtering out variants with a MAF < 0.01, INFO score < 0.8, SNPs in controls with $p < 10^{-15}$ in a Hardy-Weinberg Equilibrium Fisher's exact test (which are more likely affected by genotyping error or the effects of natural selection); SNPs that are missing in a high fraction of subjects ($> 10\%$), since this may indicate problems in the DNA sample or processing.

2. Then, NEURAL SIGNIFICANT ENHANCER-based SNPs (see section 2.4.1) were extracted from the QCed UKBB cohort.

3. Associations for enhancer-based SNPs were calculated with the schizophrenia phenotype using PLINK2 (Purcell & Chang, 2022) with flags $--glm\ --ci\ 0.95$ for the additive (standard) model, $--glm\ recessive\ --ci\ 0.95$ and $--glm\ dominant\ --ci\ 0.95$ for the alternative *recessive* and *dominant* models. The covariates included age, sex, and the first 10 PCA components of the genotypes for the UKBB population.

A plot summarising the top associated SNPs for each model was produced.

4.2.2 Internal and external validation

To validate the EP-WAS, I proceeded to calculate partitioned PRSs for each partition – the *enhancer-based* partition based on the UK Biobank-based EP-WAS, using dominant and recessive inheritance models, and the *residual* based on the standard PGC GWAS. I then computed the coefficients of determination for each partition separately, and for the composite models, as in Chapter 3. Each step is detailed below – and repeated two times, for a *dominant* and *recessive* inheritance model:

1. **Base partitioning and clumping**: this step entails creating GWAS partitions based on whether each SNP falls within a NEURAL SIGNIFICANT ENHANCER or not. The difference with the analogous step described in paragraph 3.3.1.2 is that the *enhancer-based* partition this time is based on the EP-WAS (NEURAL SIGNIFICANT ENHANCERS-based SNPs, with association measures calculated within the UK Biobank cohort using a specific inheritance model); while the *residual* partition (all non-enhancer-based SNPs) is still based on association measures from the PGC GWAS using an *additive* model. The two lists are then clumped together, so that a global PRS can be calculated without the risk of including multiple SNPs per LD block. However, to prioritise enhancer-residing SNPs, enhancer-based SNP p-values are divided by 100,000 before clumping, so that they are retained with priority over nearby residual SNPs. The specific PLINK settings are the same as those described in section 3.3.1.2, and the code can be found in this chapter's git repositories (see links in section 4.2.3).

2. **PRS calculation**: Partitioned PRSs are then calculated for each of the *enhancer-based* and *residual* partitions, using PRSice. The same settings as in section 3.3.1.4 were used in this pipeline.

3. **Calculation of the proportion of variance explained by the genetic factor – or coefficients of determination**: for each partition, CoDs were calculated using equation 3.5, and the rationale of these calculations is again described in detail in section 3.3.2. Once

more, the Nagelkerke pseudo-R^2 is also shown for comparison, but liable to sampling and other biases.

Plots summarising the total variance explained by the genetic factor on the liability scale, corrected for ascertainment (called coefficients of determination) for each partition were produced, and will be presented as Results.

4.2.3 Software and code availability

The same software was used as in section 3.3.3.

All code for this work is available on GitHub repositories:

- https://github.com/emosyne/UKBB_OR_develop for EP-WAS development, and its internal validation, both on the UK Biobank dataset.
- https://github.com/emosyne/lisa_validation for EP-WAS external validation on a PGC sample.

4.3 Results

In this chapter I performed an EP-WAS (*enhancer-based* GWAS): in other words I measured associations with schizophrenia of SNPs falling within NEURAL SIGNIFICANT ENHANCERS (please see Chapter 2 for a rationale and methods of developing *enhancer-based* partitions); to do so, I used both *dominant*, *recessive*, and *additive* inheritance models. The EP-WAS associations were calculated in UK Biobank, where they were initially internally validated, before being externally validated in a PGC cohort – namely *xs234* (please see section 3.2.1 for a description of each PGC cohort, and for the rationale for selecting *xs234* as the main one for this work).

4.3.1 Comparing Dominant and Recessive enhancer-based effects within UK Biobank

Here I measured the ORs for schizophrenia for 26,607 SNPs (this is the number before clumping) falling within NEURAL SIGNIFICANT ENHANCERS in 1,169 patients with schizophrenia and 409,710 controls of European heritage within UK Biobank. As detailed in the methods, the SNPs were selected and QCed using standard criteria, then an association OR for schizophrenia for each was calculated utilising the three main inheritance models: additive (standard, for comparison), dominant, and recessive. Following joint clumping of the *enhancer-based* and *residual* partitions, NEURAL SIGNIFICANT ENHANCER-based SNPs reduced to 13,953 for the dominant, and to 14,019 for the recessive inheritance models.

Figure 4.1 shows the odds ratios for schizophrenia for top NEURAL SIGNIFICANT ENHANCER-based SNPs in the *UK Biobank* cohort, using a **dominant** model (SNPs with dominant p-value < 0.001). Please note that, due to the number of SNPs tested ($\sim 27K$), only SNPs with p-values $< 2 \times 10^{-6}$ would be considered significant genome-wide, using a FDR of 0.05. However, as I will be using these results as part of PRS building, and will therefore consider specific thresholds as part of that work, I am showing SNPs with dominant p-value $< 10^{-3}$ as an example of the most significant results found, with no expectation that these results would be considered significant genome-wide.

Figure 4.1 shows that using a **dominant** model (**blue** dots and 95% confidence intervals) produces results that are different from those calculated using an **additive** model (**red** dots and 95% confidence intervals), specifically at rarer alleles. In fact, the SNPs towards the bottom of the plot are all showing MAFs $\sim 1 - 3\%$, at the bottom end of the spectrum for SNPs considered in GWAS (let's remember that SNPs with MAF $< 1\%$ were excluded from the analysis). One could also notice that these rarer SNPs effects, calculated using a **dominant** model in UK Biobank – based on 1,169 patients and 409,710 controls – are more extreme than those calculated in the same cohort using an additive model, and also much more extreme than those obtained meta-analysing the 90 international cohorts

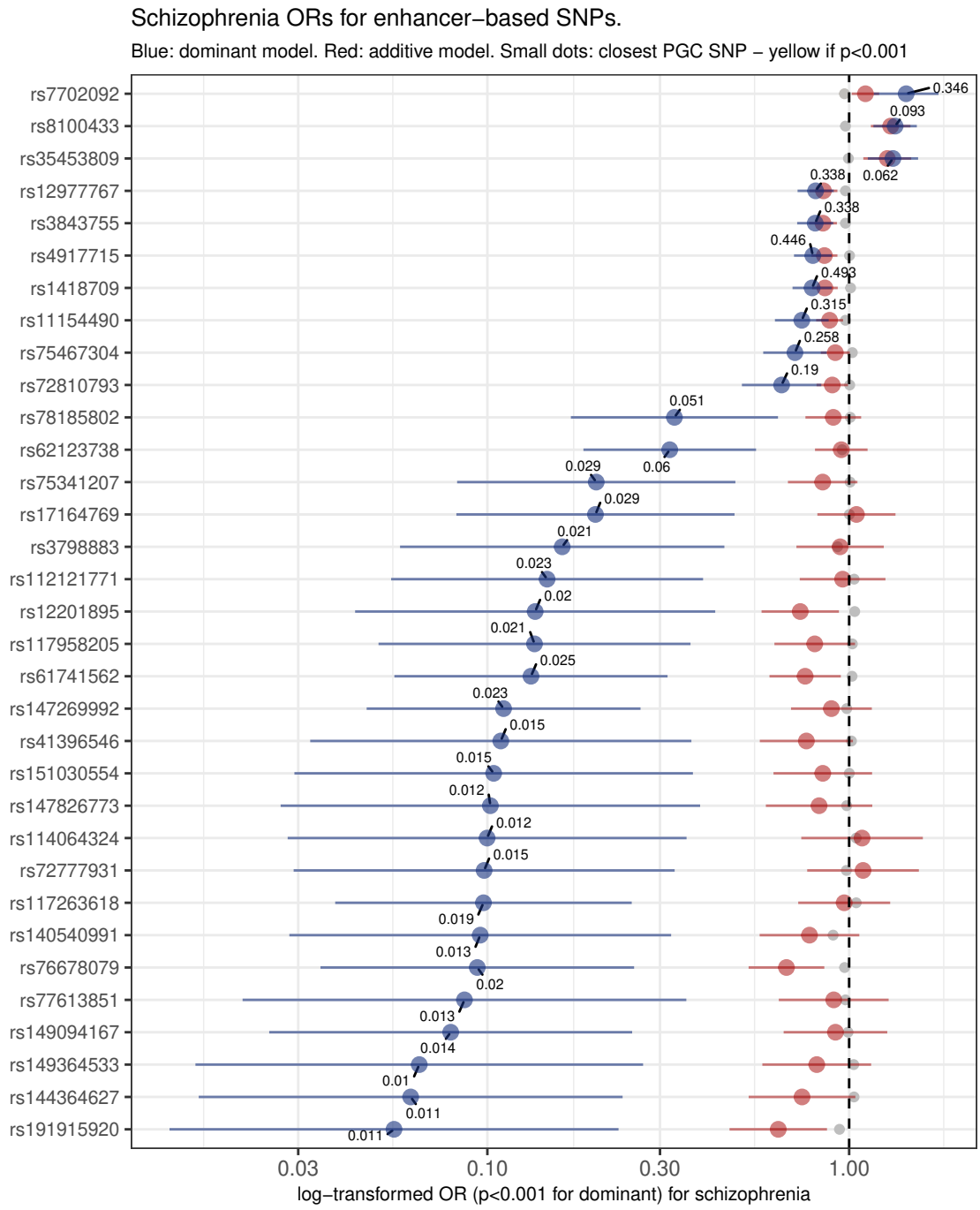


Figure 4.1: Odds ratios for schizophrenia for top *enhancer-based* SNPs, in the *UK Biobank* cohort – dominant model.
Each row in the figure represents a SNP, with its *rs...* code on the left; only top **dominant** SNPs (with dominant p-value < 0.001 in UK Biobank) are presented. **Blue** dots (and 95% confidence intervals) represent schizophrenia ORs calculated using a **dominant inheritance model**. **Red** dots (and 95% CI) represent schizophrenia ORs calculated using a standard **additive model** for comparison. The smaller **gray/yellow** dots represent schizophrenia ORs for the **closest SNP** calculated using a standard **additive model** as calculated in the PGC meta-analysis; these dots are yellow when PGC p-value < 0.001; please note: the direction of association could be reversed as the PGC SNP is not the same as the UKBB SNP, but the closest. The small text labels indicate the minor allele frequency (MAF) for each SNP, as calculated in UK Biobank; SNPs with MAF < 0.01 were excluded.

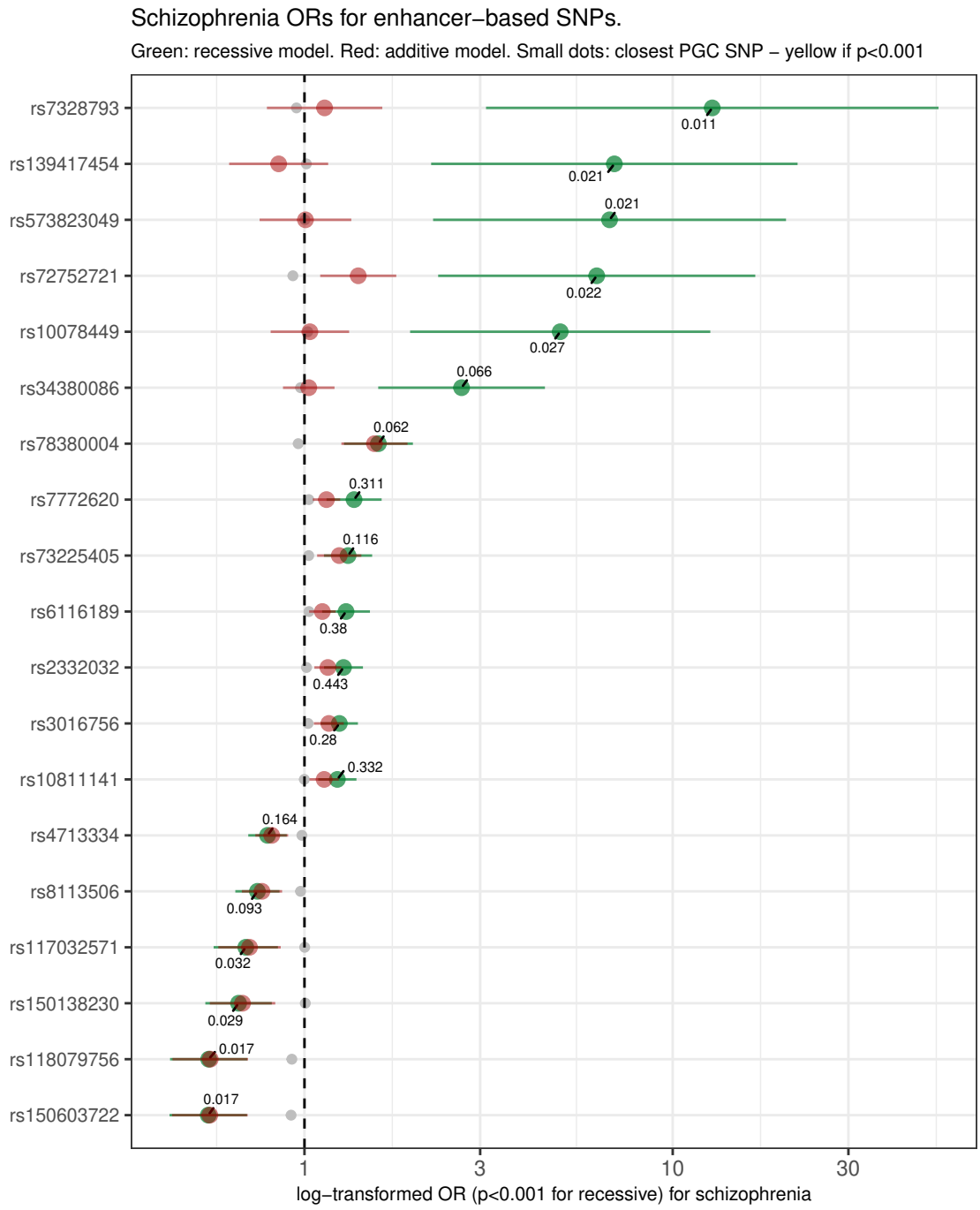


Figure 4.2: Odds ratios for schizophrenia for top *enhancer-based* SNPs, in the *UK Biobank* cohort – recessive model.
Each row in the figure represents a SNP, with its *rs...* code on the left; only top **recessive** SNPs (with recessive p-value < 0.001 in UK Biobank) are presented. **Green** dots (and 95% confidence intervals) represent schizophrenia ORs calculated using a **recessive inheritance model**. **Red** dots (and 95% confidence intervals) represent schizophrenia ORs calculated using a standard **additive model** for comparison. Smaller **gray/yellow** dots represent schizophrenia ORs for the **closest SNP** calculated using a standard **additive model** as calculated in the PGC meta-analysis; these dots are yellow when PGC p-value < 0.001; please note: the direction of association could be reversed as the PGC SNP is not the same as the UKBB SNP, but the closest. The small text labels indicate the minor allele frequency (MAF) for each SNP, as calculated in UK Biobank; SNPs with MAF < 0.01 were excluded.

included in the PGC (smaller **gray/yellow** dots, Trubetskoy et al., 2022) – based on the genotypes of 67,390 cases of schizophrenia and 94,015 control individuals. It is therefore possible to suspect that these extreme effects might be the result of the rarity of these alternative alleles, and won't be replicated in external validation.

Moving on to **recessive** effects, figure 4.2 shows the odds ratios for schizophrenia for top neural-specific, *enhancer-based* SNPs in the *UK Biobank* cohort, using a **recessive** inheritance model (top SNPs means those with a recessive p-value < 0.001 – the same caveat applies as above about lack of genome-wide significance). As evidenced in the figure, the use of a **recessive** model (**green** dots and 95% confidence intervals), produces much more extreme results than both those obtained using an **additive** inheritance model in UK Biobank (**red** dots and 95% confidence intervals), and those obtained as part of the PGC schizophrenia meta-analysis (smaller **gray/yellow** dots, Trubetskoy et al., 2022), especially for rarer SNPs.

In conclusion, the results obtained from these nonadditive models appear divergent from those obtained in the much larger PGC cohort – especially for rarer SNPs. In the next section I will investigate whether these differences might be due to 'winner's curse', or small sample bias – given that the divergence appears to be most apparent at rarer SNPs – or else if the results represent an interesting new lead. To do so, I will first validate the results in the same UK Biobank cohort where these were developed. Then, I will externally validate them on a separate, external cohort from the PGC consortium.

4.3.2 EP-WAS internal validation in UK Biobank

To validate the EP-WAS, I calculated partitioned PRSs for each partition – the *neural-specific, enhancer-based* partition, this time based on the UK Biobank EP-WAS, using both a dominant and recessive inheritance models – and the *residual* and *original GWAS* partitions, both based on the standard PGC GWAS and using an additive model. Then, using the UK Biobank as the target population, I calculated the coefficients of determination for each partition, and for the multivariable models.

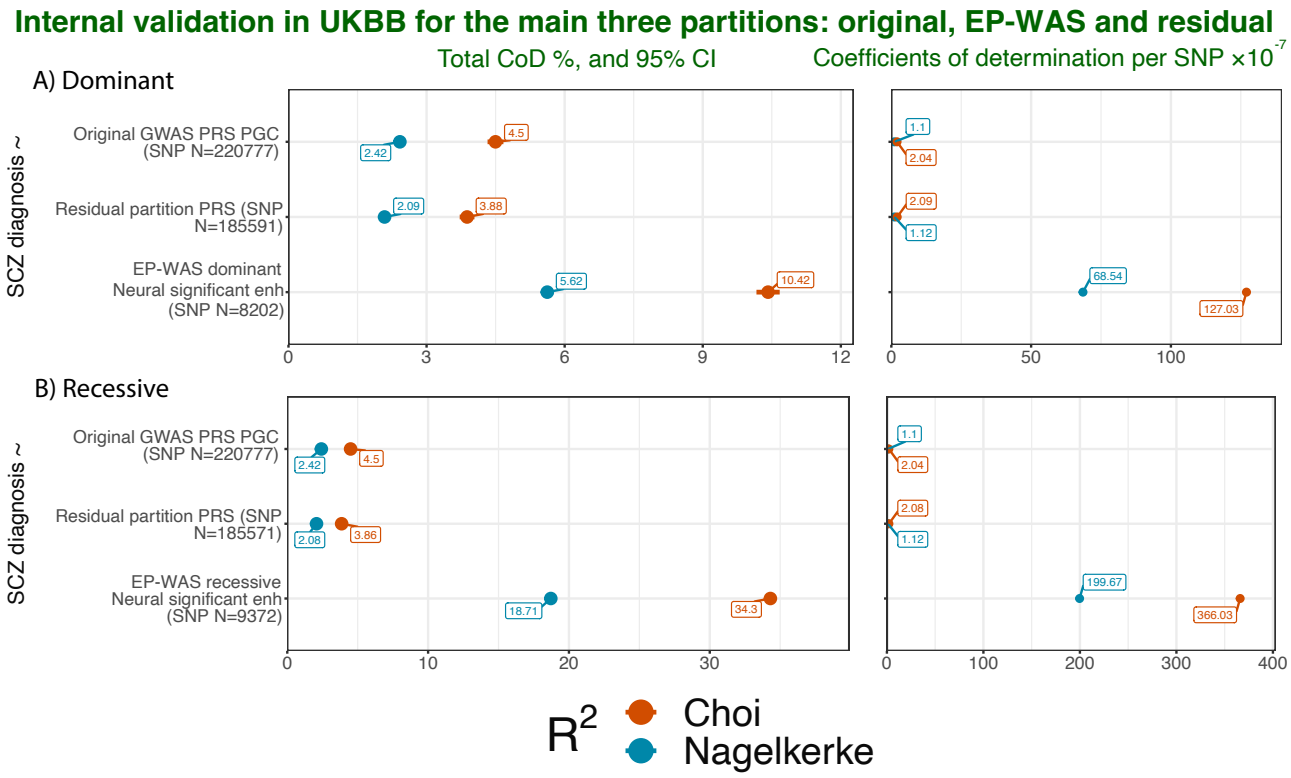


Figure 4.3: Coefficients of determination for schizophrenia for the three main partitions – original GWAS, EP-WAS for NEURAL SIGNIFICANT ENHANCERS, and *residual* – EP-WAS internal validation in the *UKBB* cohort.

The figure describes the proportion of schizophrenia variance explained by the genetic factor for the three main genomic partitions – original GWAS, *residual*, and EP-WAS *neural-specific* enhancers, in the *UKBB* cohort. The EP-WAS is based on two different inheritance models: dominant (**Panel A**) and recessive (**Panel B**). In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Each plot on the left shows the overall CoD % and 95% confidence interval, and on the right the corresponding point value adjusted per SNP ($\times 10^{-7}$).

Base data: European ancestry PGC GWAS for schizophrenia (Trubetskoy et al., 2022) for the Original GWAS and residual partitions; UK Biobank-based EP-WAS for the enhancer-based partition. Target data: UK Biobank European ancestry cohort.

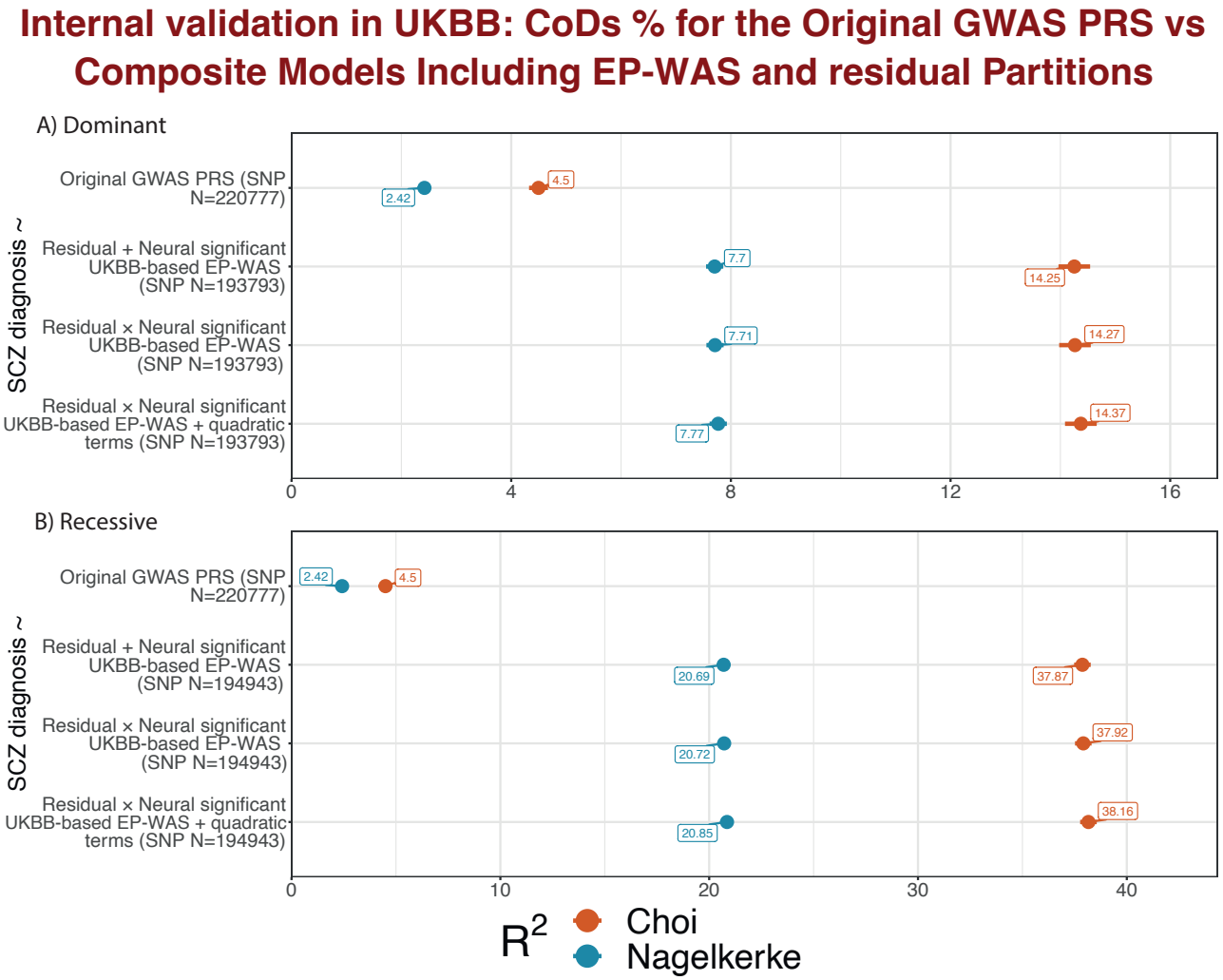


Figure 4.4: Coefficients of determination for schizophrenia for the original GWAS PRS vs composite models including EP-WAS for NEURAL SIGNIFICANT ENHANCERS – EP-WAS internal validation in the *UKBB* cohort.

The figure describes the proportion of the variance of schizophrenia explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $SCZ \sim EP\text{–}WAS\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $SCZ \sim EP\text{–}WAS\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $SCZ \sim EP\text{–}WAS\_PRS \times residual\_PRS + EP\text{–}WAS\_PRS^2 + residual\_PRS^2$.

The EP-WAS in each panel is based on two different inheritance models: dominant (**Panel A**) and recessive (**Panel B**).

Base data: European ancestry PGC GWAS for schizophrenia (Trubetskoy et al., 2022) for the Original GWAS and residual partitions; UK Biobank-based EP-WAS for the enhancer-based partition. Target data: UK Biobank European ancestry cohort.

Figure 4.3 shows that – as expected in internal validation – the CoDs for the UKBB-based EP-WAS partitions for both inheritance models were very high within UK Biobank. The **dominant** model, as applied to the NEURAL SIGNIFICANT partition, yielded a CoD of 10.42%, while the **recessive** model's CoD equated to 34.3%. These were both much higher than the 3.72% of adjusted variance explained by the NEURAL SIGNIFICANT partition pPRS based on PGC effects (see section 3.4, and Figure 3.3). Interestingly, the **recessive** model seemed to explain more variance at internal validation, with a CoD per SNP ~ 3 times than that of the **dominant**.

Similar results were apparent when using these UKBB-based EP-WAS partitions in multivariable models. Figure 4.4 shows that, while the original PGC-based PRS had a schizophrenia CoD of just 4.5% in UK Biobank, models combining the residual partition (also PGC-based) with UKBB-developed EP-WAS partitions showed a large CoD boost. More specifically, for a **dominant** model, the CoD reached 14.25% when considering a simple additive, and 14.37% when using the fully interactive model including quadratic terms. For a **recessive** model, the CoD reached 37.87% when considering a simple additive, and 38.16% when using the fully interactive model including quadratic terms.

In the next section I will explore if these encouraging results from internal validation replicate when externally validated in a separate target sample.

4.3.3 EP-WAS external validation in a PGC cohort

In this section I have again validated the EP-WAS, utilising the same methodology as in the previous section – this time on an external PGC cohort. This was necessary as the EP-WAS was developed on the UK Biobank, and therefore the coefficients of determinations generated on UK Biobank as a target population appeared inflated. Therefore, here I have calculated partitioned PRSs for each partition – the *enhancer-based* partition based on UKBB *dominant* and *recessive* EP-WASes – and the *residual* and *original GWAS* partitions, both based on the *xs234* leave-one-out (LOO) PGC original *additive* GWAS. Then, using the *xs234* PGC cohort as the target population, I calculated the coefficients of determination (CoDs) for each

partition, and for the multivariable models.

From examining Figure 4.5, it is evident that, while the LOO GWAS overall CoD for schizophrenia in this cohort was once more equivalent to 9.85%, the NEURAL-SPECIFIC ENHANCER-based pPRS for the UKBB-based EP-WAS was equivalent to just 0.04% for the *dominant*, and to 0.01% for the *recessive* models, as compared to 3.72% for the same NEURAL SIGNIFICANT ENHANCERS partition pPRS based on PGC effects (see figure 3.3 in section 3.4). As a consequence, the CoD per SNP for UKBB-based EP-WAS pPRSs was also close to null, and it would be meaningless to compare the results between the dominant and recessive models given the modest results.

Similar (negative) results, as expected, are reflected in multivariable models. Figure 4.6 shows that all models combining a *residual* partition (based on additive effects on the PGC LOO GWAS) with the UKBB-based EP-WAS pPRSs – both utilising dominant or recessive inheritance models – showed a smaller CoD for schizophrenia than the original PGC GWAS.

4.4 Summary of findings

In this chapter I have explored the potential added value of utilising *dominant* vs *recessive*, as compared to canonical *additive* SNP inheritance models, for a schizophrenia GWAS. To do so, I first calculated the ORs for schizophrenia for enhancer-based SNPs within the UK Biobank cohort using all three these inheritance models – something I named an EP-WAS, and compared the results for the top SNPs for each model. I found that, while for common SNPs the models did not seem to show disparate results, for rarer SNPs the ORs became more extreme for both a *dominant* and a *recessive* models, as compared to the *additive* one.

I then validated the results in terms of the proportion of the total variance explained by PRS on the liability scale, corrected for ascertainment. I first calculated this measure for each genomic partition (original PGC GWAS, NEURAL SIGNIFICANT ENHANCERS,

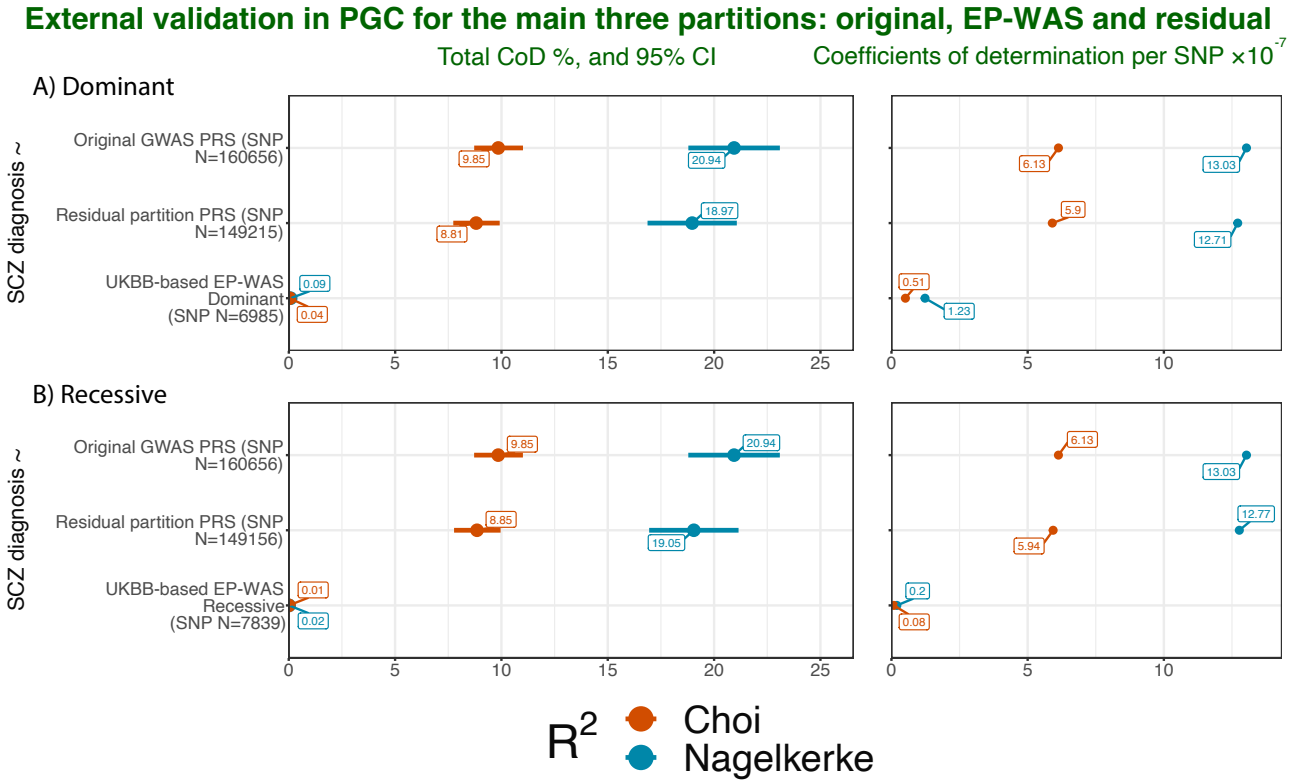


Figure 4.5: Coefficients of determination for schizophrenia for the three main partitions – original GWAS, EP-WAS for NEURAL SIGNIFICANT ENHANCERS, and *residual* – EP-WAS external validation in the *xs234* PGC cohort.

The figure describes the proportion of schizophrenia variance explained by the genetic factor for the three main genomic partitions – original GWAS, *residual*, and EP-WAS *neural-specific* enhancers, in the *UKBB* cohort. The EP-WAS is based on two different inheritance models: dominant (**Panel A**) and recessive (**Panel B**).

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Each plot on the left shows the overall CoD % and 95% confidence interval, and on the right the corresponding point value adjusted per SNP ($\times 10^{-7}$).

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort for the Original GWAS and residual partitions; UK Biobank-based EP-WAS for the enhancer-based partition. Target data: xs234 European PGC schizophrenia cohort.

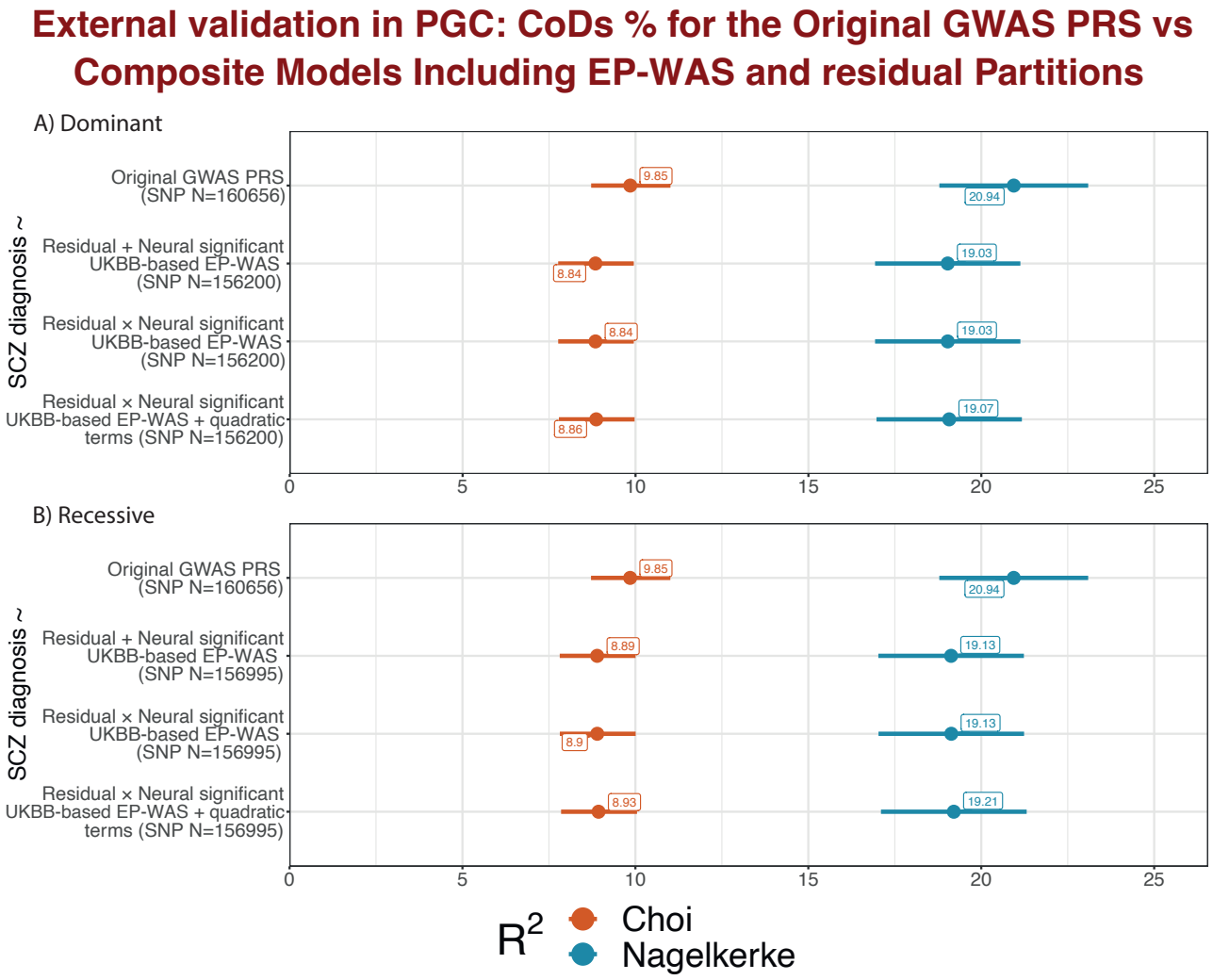


Figure 4.6: Coefficients of determination for schizophrenia for the original GWAS PRS vs composite models including EP-WAS for NEURAL SIGNIFICANT ENHANCERS – EP-WAS external validation in the *xs234* PGC cohort.

The figure describes the proportion of the variance of schizophrenia explained by the genetic factor. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoDs) – and 95% confidence intervals. In baby blue the original Nagelkerke's R^2 for comparison. In each panel are represented the CoDs, from top to bottom for: ① The original GWAS PRS, for comparison; ② Logistic model 1 – simple additive: $SCZ \sim EP\text{–}WAS\_PRS + residual\_PRS$; ③ Logistic model 2 – additive model plus interactions: $SCZ \sim EP\text{–}WAS\_PRS \times residual\_PRS$; ④ Logistic model 3 – additive model + interactions + quadratic terms: $SCZ \sim EP\text{–}WAS\_PRS \times residual\_PRS + EP\text{–}WAS\_PRS^2 + residual\_PRS^2$.

The EP-WAS in each panel is based on two different inheritance models: dominant (**Panel A**) and recessive (**Panel B**).

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort for the Original GWAS and residual partitions; UK Biobank-based EP-WAS for the enhancer-based partition. Target data: xs234 European PGC schizophrenia cohort.

and *residual* partitions) within the UKBB development sample: this showed that *enhancer-based* SNP effects from the EP-WAS were very large, and much larger than expected. The results within UK Biobank, however, appeared inflated, due to the fact that the EP-WAS was also developed in UKBB. Finally, I validated the results externally in the *xs234* PGC cohort. This external validation step showed that EP-WAS effects did in fact suffer from a large 'winner's curse' bias, as these results based on an external cohort not only did not improve on the original PGC-based effects, but they showed a very significant drop to the amount of variance explained by PRS. The findings are discussed in Chapter 5.

Chapter 5

Discussion

Since the mid-2000s (DeWan et al., 2006; Klein et al., 2005; The Wellcome Trust Case Control Consortium, 2007) we have witnessed a sharp increase in the number of GWASes. Due to the need for ever larger development samples, the field of psychiatry has seen the formation of large consortia, such as the Psychiatric Genomics Consortium (PGC, https://pgc.unc.edu/). The PGC Schizophrenia working group, for example, is a group of researchers and clinicians world-wide who share the common aim to advance the genomics of schizophrenia; they have worked by collecting samples that accumulate over time, and publishing the Consortium papers in *waves*. Together with discovering new loci, the increase in sample size at each wave has led to a very welcome increase in the replicability of genetic findings: as an example, in the latest PGC schizophrenia GWAS (wave 3, including over 53K cases, Trubetskoy et al., 2022), all but one of the 108 genome-wide hits from

the previous wave 2 analysis (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) were replicated. However, the amount of variance explained by PRS ($P_T = 0.05$) on the liability scale has failed to keep up with the increases: this has gone from 7.0% in PGC wave 2 to 7.3% in PGC wave 3, using all ancestries (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Trubetskoy et al., 2022). This is well below the estimated heritability of $\sim 64 - 80\%$ for schizophrenia (Lichtenstein et al., 2009; Sullivan et al., 2003).

In this thesis I decided to tackle the issue of this 'missing heritability' (see section 1.5.2) through the study of two model conditions: schizophrenia and HCM. I have picked schizophrenia as the main model condition due to its prevalence, unmet therapeutic need, high genetic burden, and neurodevelopmental nature. The genetic susceptibility and environmental risk factors for schizophrenia are likely to converge on the neuron, and its genetic risk from GWAS is enriched in non-coding areas. Further, schizophrenia genetic risk variants have been shown to preferentially fall within regions containing ultra-conserved genetic elements and GRBs, as discussed in more detail in section 1.6.1 and in Barešić et al., 2020. I have then decided to use hypertrophic cardiomyopathy (HCM) as a comparison to schizophrenia, as discussed in section 1.6.2, due to its different genetic architecture, likely more pronounced tissue-specific genetic milieu, and lower likelihood of a developmental aetiology. HCM has several high-risk variants and is less polygenic than schizophrenia (Mazzarotto et al., 2020). Most high-risk variants affect the sarcomere within the heart, and the disease mostly appears in adults.

In this work I have addressed the overarching issue of complex disorder 'missing heritability' from several perspectives:

I. In **Chapter 2,** ***Schizophrenia and HCM heritability enrichment in tissue-specific enhancers***, I began by measuring the amount of heritability for schizophrenia and HCM that resides within various regulatory genomic partitions including several classes of enhancers, and whether heritability levels diverged from the expected. This work allowed me to select heritability-enriched partitions to use for subsequent analysis.

II. In **Chapter 3, *Schizophrenia and HCM heritability from partitioned PRSs***, I strived to address the same issue of 'missing heritability' by prioritising enhancer-based SNPs as follows:

• First I developed **partitioned polygenic risk scores, or pPRSs** – these are PRSs where SNPs based in different parts of the genome (e.g., within tissue-expressed enhancers or not) are separated. pPRSs allowed me to study the contribution to complex disorder heritability of *tissue-specific, enhancer-based* partitions separately, and with priority over, the partition representing non-enhancer-based SNPs, which I have called *residual*.

• Then, through the use of **multivariable models** including both *enhancer-based* and *residual* pPRSs, I explored the adjusted PRS-based variance explained by each predictor separately, as well as measuring the contribution of interaction and quadratic terms.

• Finally, for enhancer-based SNPs I multiplied SNP-disease association measure β coefficient by either the *effect size* of each tissue-specific enhancer (its association measure with a target gene), or by its tissue-specific expression, to test if enhancer-based annotations could improve overall disease h^2_{pPRS}.

III. In **Chapter 4, *Leveraging nonadditive disease inheritance models***, I explored the h^2_{pPRS} contribution of alternative inheritance models (i.e., additive or recessive) – through the development of an enhancer-based schizophrenia association study (EP-WAS). I then validated the findings both internally within UK Biobank, and externally in a separate schizophrenia cohort taken from the PGC.

I will discuss each chapter's findings, as well as how they fit in with my initial hypotheses, and their potential limitations, below.

5.1 Chapter 2: Schizophrenia and HCM heritability enrichment in tissue-specific enhancers

Chapter 2 builds upon previous work from the Lenhard and Howes labs (Barešić et al., 2020), which had investigated the potential role of genomic regulatory blocks (GRBs) in understanding the genetics of neuropsychiatric disorders. The authors had reviewed GRB-based approaches to assigning loci in non-coding regions to potential target genes and had applied them to reanalysing the results of the (then) largest schizophrenia GWAS (PGC wave 2, Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). The authors had found that disease-associated SNPs are over-represented in GRBs, and that the GRB model is a powerful tool for linking these SNPs to their correct target genes under long-range regulation.

In this chapter I took this approach further, by looking at all bidirectionally transcribed enhancers identified in FANTOM5 genome-wide (Andersson et al., 2014), and not just at those within GRBs. Secondly, working with the AR+C dataset (Georgieva, 2022), as well as with the FANTOM5 resource, I created genome-wide lists of neural- and heart-expressed enhancers, with evidence of co-expression and chromatin 3D contact with at least one gene within 3Mb (significant, tissue-specific enhancers). Finally, I tested whether these tissue-specific enhancer lists (as well as a number of control lists) showed enrichment in the heritability for exemplar conditions schizophrenia and HCM, updating the findings to the latest GWASes to date (schizophrenia PGC wave 3 by Trubetskoy et al., 2022, and HCM Consortium by Tadros et al., 2023).

Due to schizophrenia's highly polygenic genetic architecture, as well as due to its neuro-developmental credentials, my hypothesis was that neural-tissue-expressed enhancer genomic compartments would be enriched in schizophrenia-associated common genetic variants. On the other hand, due to HCM's less polygenic architecture, as well as its likely more tissue-specific genetic contribution, I had hypothesised that cardiac-tissue-expressed enhancer genomic compartments would not be enriched in HCM-associated common ge-

netic variants.

As part of the results, I found confirmation that neural-expressed enhancers were highly enriched in schizophrenia heritability, and that this enrichment was specific to those enhancers falling within human-mouse GRBs. On the other hand, heart-expressed enhancers appeared enriched in HCM heritability, however not to a statistically significant extent after adjusting *p*-values for multiple testing. These findings were largely in accordance with my expectations and hypothesis, and extend the literature on this topic, including confirming previous findings from Barešić et al., 2020, as discussed above.

5.1.1 Interpretation of findings

Based on the existing literature, my first hypothesis (see section 1.7) was precisely that GRBs and long-range enhancer-based regulatory elements would be enriched in schizophrenia heritability, while HCM would not. Both these hypotheses proved correct. Why was I able to anticipate these findings? First of all, as described in section 1.2.2 in the introduction, as well as in section 2.1.1, conserved genetic elements such as GRBs have been studied in the context of the regulation of human (and animal) development (Polychronopoulos et al., 2017). In particular, CNEs have been studied as cis-regulatory elements coordinating spatial-temporal gene expression, especially during embryonic development (Sandelin et al., 2004).

Schizophrenia, as we have seen in section 1.6.1, is a neuro-developmental condition which is likely to owe its manifestation in early to mid life to multiple insults, usually considered to be a combination of genetic, pre-birth, early life as well as later life insults (McCutcheon et al., 2019). What most, if not all, theories about the development of schizophrenia have in common, is the focus on the fact that there seems to be a convergence on the synapse and on abnormal neuron-neuron communication (Howes & Onwordi, 2023; McCutcheon et al., 2020; Onwordi et al., 2020; Osimo et al., 2019). Multiple lines of evidence have linked GRBs, as well as to development in general, as we have seen, more specifically to neuro-development, and the formation and maintenance of synapses: for example, Geor-

gieva, 2022 recently found that predicted GRB target genes were enriched in gene ontologies including *axon development*, *embryonic organ development*, *forebrain development*, *axon guidance*, and *neuron projection*, among others. These findings, surprisingly, were based on genome-wide analyses of enhancer-based GRB target genes, without even the need to subset neural-expressed enhancers. Furthermore, Barešić et al., 2020 had already found an enrichment in schizophrenia signal in human-mouse GRBs. They had further concluded that, because of the known involvement of GRB target genes in regulation of development, the GRB model might represent a powerful tool for linking schizophrenia SNPs to their correct target genes under long-range regulation.

The fact that cardiac-tissue-specific enhancer lists, as well as GRBs, did not appear significantly enriched in HCM heritability can be interpreted as a potential sign that HCM might be a condition that develops because of stronger, more targeted, molecular signals, which might affect a terminally differentiated tissue such as the myocardium. This would be compatible with the genetic aetiology of HCM, a condition showing a polygenic inheritance but with a stronger component of high-risk mutations, which are known to be directly affecting the sarcomere (Mazzarotto et al., 2020).

In conclusion, we know that both HCM and schizophrenia are *complex* genetic disorders, as defined and discussed in section 1.4.1.1, however they show varying degrees of polygenicity, with schizophrenia notoriously at the higher end of the polygenicity spectrum (Visscher et al., 2021), and HCM likely at the other, lower end; moreover, schizophrenia-related genes are known to rely on a wider network of regulatory elements, which justifies the higher number of non-coding associated regulatory variants (Roussos et al., 2014). Furthermore, my findings support the developmental (and even more specifically, *neuro*-developmental) credentials of enhancer-based gene regulation.

5.1.2 Limitations

The main limitation of this work is the use of FANTOM5 data (Andersson et al., 2014), as further analysed in the AR+C (Georgieva, 2022). While the human section of the

dataset, which I have used, is very rich, and includes annotations on co-expression based on CAGE data on 800 human samples – 241 cell line, 447 primary cell, and 120 tissue samples – the preponderance of cell lines and primary cells over tissue samples might have meant that the results would be limited in their ability to represent the complexity of the human brain – or the human heart. However:

- Georgieva, 2022 has already benchmarked her results against existing enhancer-target gene assignment methods, as discussed in section 1.3.1, and shown that the AR+C excels in differentiating long-distance enhancer-promoter regulatory interactions in developmental genes, which overlaps with the aims of this work.
- AR+C co-expressed enhancer-promoter pair predictions were further refined based on 3D chromatin interaction data. The Hi-C and Micro-C datasets used for this aim included both cardiac- (Zhang et al., 2019) and neural-specific (Akbarian et al., 2015) human interaction data, as further introduced in section 2.2.1.

Further, to overcome this limitation, I compared AR+C-based results with the PsychENCODE pre-frontal cortex 'high-confidence' enhancers dataset as a sensitivity analysis: LDSC on this dataset showed a high degree of overlap with AR+C-based results – including enrichment in schizophrenia heritability; however, the NEURAL SIGNIFICANT ENHANCERS WITHIN GRBS AR+C-based list, generated as part of this work, showed an even higher enrichment in heritability (see figure 2.2), suggesting that AR+C-based lists compensate for the potential limitations due to the use of cells with the high quality of available annotations, based on co-expression, 3D contact, as well as data on tissue-specific expression.

5.1.3 Further work

This chapter – by developing tissue-specific enhancer lists that are enriched in schizophrenia heritability – formed the basis for the subsequent analyses, presented in chapters 3 and 4 – discussed next. In these chapters I have utilised the tissue-specific enhancer lists generated here, aiming to increase the amount of disease heritability explained by schizo-

phrenia PRSs.

5.2 Chapter 3: Schizophrenia and HCM heritability from partitioned PRSs

In Chapter 3, **Schizophrenia and HCM heritability from partitioned PRSs**, I developed **partitioned PRSs**, or **pPRSs**, to dissect the heritability contribution of tissue-specific enhancer-based SNPs separately, and with priority over, every other SNP. Because of the heritability enrichment in tissue-specific enhancers for schizophrenia, which I showed in Chapter 2, I had reason to believe that enhancer-residing SNPs would carry a higher schizophrenia per-SNP h^2_{SNP}. Therefore, I had hypothesised that the h^2_{pPRS} for schizophrenia would increase by considering **prioritised** *tissue-specific enhancer*-based SNPs and *residual* partitions separately, as compared to the original GWASes for the condition. Given the non-significant enrichment in heritability for HCM for cardiac-specific enhancers, I hypothesised the same for HCM, however with a lower confidence.

Confirming the results of Chapter 2, I found that – for schizophrenia – neural-specific enhancer-based SNPs, and especially those within human-mouse GRBs, explained several times the per-SNP variance than the average GWAS SNP; this was also the case for cardiac-specific enhancers and HCM. Further, using logistic models where the tissue-specific, *enhancer-based* and *residual* partition PRSs were separate predictors, I tested whether the total variance explained by the genetic factor on the liability scale, corrected for ascertainment – called coefficient of determination – by pPRS (h^2_{pPRS}) would increase over that explained by the original GWAS for schizophrenia and for HCM (h^2_{PRS}). For schizophrenia the use of multivariable models produced only modest increases compared to the h^2_{PRS} figure – and with quite wide confidence intervals, which does not allow to refute the null hypothesis. The best multivariable model, for the *logit* additive model plus interactions plus quadratic terms including NEURAL SIGNIFICANT ENHANCERS WITHIN GRBS, reached a CoD of 10.49%, which, compared to the original leave-one-out GWAS coefficient of 9.85%,

represents a 6.5% improvement.

Similarly, for HCM the use of multivariable models seemed to drive an increase in h^2_{pPRS}, as compared to the original GWAS PRS alone (h^2_{PRS}), in both the UK Biobank and the Royal Brompton cohorts – even if the increases for the UK Biobank cohort were much larger. Differently from schizophrenia, however, this increasing trend was apparent both for the CARDIAC SIGNIFICANT ENHANCERS partition, as well as for non associated partitions (derived from 'control' enhancer lists, where the enhancers did not necessarily have any association with a promoter, and/or were not necessarily cardiac-specific), suggesting that this pattern might not be driven by the use of tissue-specific enhancer-based partitions. In fact, the main improvement in h^2_{pPRS} came with the use of quadratic terms. This appeared to suggest that up-weighting extreme PRS values – in this case by the use of a quadratic PRS term – might benefit less polygenic, and more tissue-specific conditions such as HCM.

Finally, in this chapter I have tested if enhancer-based annotations can help improve PRSs. The rationale for this analysis resides in the importance of enhancer-promoter interactions for tissue- and time-specific gene expression regulation (as introduced, among others, in section 1.1.3.2). Previous research had also shown that GRBs can act as regulatory domains that delimit the span of long-range gene regulatory interactions (section 1.2.2, and Barešić et al., 2020; Georgieva, 2022). Therefore, it is possible to hypothesise that the genetic effects of a SNP tagging a specific enhancer might be affected by either how strong the enhancer-promoter association is for the same enhancer, or by how expressed the enhancer is, in the specific tissue of interest. For these reasons, in this analysis I have included two measures: the association *effect size* for enhancer-promoter co-expression from the AR+C – a measure of how strong an enhancer can modulate its target gene – as well as a measure of enhancer tissue-specific expression (neural or cardiac for each analysis). In other words, I tested the hypothesis that tissue-specific enhancer PRSs for schizophrenia and HCM would increase the overall disease h^2_{pPRS} when accounting for tissue-specific enhancer expression or target gene association measures – by multiplying SNP-disease association measure β coefficient for enhancer-based SNPs by either the *effect size* of the tissue-specific enhancer,

or by its tissue-specific expression. This part of the work did not show any promise, as the results were negative for both conditions.

5.2.1 Interpretation of findings

This chapter moved from the premises set in my hypothesis that the h^2_{pPRS} for schizophrenia and HCM would increase by considering *tissue-specific enhancers* and *residual* partitions separately, as compared to the original GWASes for the conditions, and that the *tissue-specific* enhancers partition PRSs for schizophrenia and HCM would increase by accounting for tissue-specific enhancer expression or target gene association measures. As discussed more extensively elsewhere, these hypotheses were driven by multiple sources of previous evidence, and in summary:

- GWAS, paired to polygenic risk scoring, are techniques that allow to estimate SNP-based heritability for *complex* disorders. However, h^2_{PRS} estimates have been shown to be consistently smaller than disease h^2, as measured by twin studies (see section 1.5.2).

- GWAS does not take into account tissue-specific annotations or expression features, nor it does account for the theories around what might drive a specific condition's heritability, such as the fact, for example, schizophrenia heritability has been shown to culminate in neural-specific effects, and particularly on the synapse (see section 2.1).

- This work makes use of annotations based on co-expression and 3D chromatin contact information that allowed to create a genome-wide list of high-confidence enhancer-promoter pairs. These were refined with tissue-specific enhancer expression information to generate tissue-specific enhancer lists (see section 2.2).

Based on previous knowledge on GRB biology, which assigns a role to GRBs in long-range developmental gene regulation, tissue-specific enhancer lists were further subset based on GRB overlap (see section 1.2.2).

- Work from Chapter 2 in this thesis, that confirms that tissue-specific enhancer lists, as well as GRBs, appear to be enriched in schizophrenia heritability.

This chapter's results appear conclusive in one main respect: in showing that

tissue-specific enhancer-based partitions appear to carry significantly more heritability (between 8 and 13 times) than non-enhancer-based ones for schizophrenia, and also for HCM (between 5 and 19 times). The other result is to show that the use of multivariable models seems to drive a small increase in h^2_{pPRS} for both schizophrenia and HCM, as compared to the original GWAS PRS alone (h^2_{PRS}). However, the fact that for HCM this pattern was apparent both using a significant cardiac enhancers partition, as well as non associated partitions (where the enhancers did not necessarily have any association with a promoter, and were not cardiac-specific), suggests that this pattern might not be necessarily driven by the use of tissue-specific enhancer-based partitions. In fact, the main improvement in h^2_{pPRS} came with the use of quadratic terms. This appears to suggest that up-weighting extreme PRS values – in this case by the use of a quadratic PRS term – might benefit less polygenic, and more tissue-specific conditions such as HCM.

In conclusion, I am unable to reject the null for both my main hypotheses for this chapter, with relation to schizophrenia: NEURAL SIGNIFICANT ENHANCER-based partitions did not show significant improvements in h^2_{pPRS} for schizophrenia, and the same measure did not increase when accounting for tissue-specific enhancer expression or target gene association measures. Previously, regulatory annotations had been used to improve PRS trans-ancestry 'portability' (Amariuta et al., 2020; Weissbrod et al., 2022), producing modest relative improvements in the explained disease liability when applying a GWAS developed in one population to a different ethnicity. In Márquez-Luna et al., 2021, the Authors introduce *LDpred-funct*, a tool that leverages trait-specific functional priors to increase prediction accuracy. The model considers predictors including whether a variant is coding, conserved, regulatory, and LD-related annotations. Using this method, they produce a +4.6% relative improvement in R^2, which is comparable in scale to the 6.5% improvement in CoD that I find in this work. My findings, therefore, despite not being likely clinically relevant or statistically significant due to the wide confidence intervals, are comparable to previous work on the topic. It is possible that functional annotations might not grant very large improvements in the amount of variance explained by the genetic factor, and in future work it might

be best to combine existing approaches.

Finally, as with most negative findings, there is the possibility that I did not have the statistical power to detect differences in CoDs, e.g. a type II error. This might be the case especially for analyses that involved enhancer-list-based genomic partitions, as compared to *residual* partitions, as these were many times the size of the enhancer-based ones. This is especially true with regards to smaller enhancer lists, such as the lists overlapping eQTLs. In fact, as an example, BRAIN ENH-PROMOTER-EQTLS did show large values of heritability enrichment for schizophrenia, however these did not pass the FDR threshold. For this very reason I excluded these smaller E-P_eQTL lists from the analyses in this chapter.

5.2.2 Limitations

The main limitation to this work resides in how tissue-specific enhancers were selected, and then in how PRSs were partitioned. First of all, the approach relied on enhancer expression in any tissue/cell: for example, an enhancer was marked as 'neural' if it was expressed in any of a list of neural-related FANTOM5 cells or tissues. This approach, despite being not very specific, did generate lists appropriate to warrant consideration in the remainder of this work. In future work, more refined measures of enhancer tissue specificity could be utilised alongside this. However, as noted earlier, the comparison of the neural-specific lists I generated for this work using this approach with existing ones (such as the PsychENCODE pre-frontal cortex 'high-confidence' enhancers) produced favourable results, something that supports the current approach.

Secondly, the generation of enhancer-based PRSs relied on selecting SNPs within ± 100 bps of enhancer coordinates, and then prioritising them over nearby SNPs in reciprocal LD when clumping. This approach is sensitive, however it does not capture any SNPs for a number of significant enhancers – for example because no SNPs were genotyped within 100 bps of a given enhancer. The variance of such missed enhancers is therefore lost. Further, the approach prioritises **all** available SNPs within significant enhancers, without selecting enhancers that might be more or less promising. Of course, if an enhancer is within a 'signi-

ficant' list, this will mean that it has features that make it – on average – more important than nearby DNA. However, in future, approaches tailored to specific enhancers – e.g. based on some sort of score – and the use of genome sequencing to capture most enhancers, might be more fruitful.

Further, as more extensively discussed in section 5.1.1, there are limitations to the work behind the creation of the tissue-specific enhancer lists, as these are based on FANTOM5 data (Andersson et al., 2014), as further analysed in Georgieva, 2022. While this is a very rich dataset, including various human cells and tissues, its neural samples might not accurately represent the complexity of the human brain – more specifically, these annotations might not capture some of the regulatory elements (enhancers) important for gene regulation in the human brain (and relevant to schizophrenia). However, I did find significant results for enrichment in schizophrenia heritability – which makes me confident that the dataset might be effective at building relevant tissue-specific lists which are highly enriched in schizophrenia heritability.

5.2.3 Further work

I am aiming to combine existing approaches such as *LDpred-funct* (Márquez-Luna et al., 2021) with mine, to test whether the combined use of multiple annotations can further increase the amount of variance for schizophrenia explained by the genetic factor.

5.3 Chapter 4: Leveraging nonadditive disease inheritance models

Chapter 4, **Leveraging nonadditive disease inheritance models**, explored the value of using alternative, nonadditive inheritance models to developing an enhancer-based schizophrenia association study (EP-WAS). This analysis starts from the observation that A) to study each SNP's association with a trait or phenotype within a GWAS one has to make assumptions about a specific inheritance model (as discussed in section 1.4.2.1); and B) most

existing GWASes for complex disorders have utilised an *additive model* of inheritance – this assumes that the risk for each additional SNP is small, and that each additional allele acts independently by increasing risk (Psychiatric GWAS Consortium Coordinating Committee, 2009; Uffelmann et al., 2021). Further, while most Mendelian disorders and conditions are classified as following a dominant or a recessive inheritance pattern, and it is postulated that most deleterious mutations are recessive, *nonadditive* effects have seldom been studied in the context of SNP-based heritability for complex human disorders (Manna et al., 2011). On the other hand, *dominance effects* are central to the study of model disease fitness by population geneticists, and across organisms and conditions the average dominance of mutation of small effects should be approximately one-quarter (Manna et al., 2011). As a consequence, I have hypothesised that some higher-priority variants (e.g., those falling inside tissue-specific enhancers for a relevant tissue) might follow a *nonadditive* (e.g., dominant/recessive) model of inheritance, and that considering dominant or recessive inheritance models for enhancer-based SNPs in schizophrenia could increase its h^2_{pPRS}.

To test this hypothesis, in this chapter I measured the associations of enhancer-based SNPs with schizophrenia using both *dominant* and *recessive*, as compared to canonical *additive* inheritance models, within the UK Biobank cohort. I then validated the results in terms of the proportion of the total variance explained by PRS on the liability scale, corrected for ascertainment. I first calculated this measure for each genomic partition (original PGC GWAS, NEURAL SIGNIFICANT ENHANCERS, and *residual* partitions) within the UKBB development sample: this showed that *enhancer-based* SNP effects from the EP-WAS were very large, and much larger than expected. The results within UK Biobank, however, appeared inflated, due to the fact that the EP-WAS was also developed in UKBB. Finally, I validated the results externally in the *xs234* PGC cohort. This external validation step showed that EP-WAS effects within UKBB did in fact suffer from a large 'winner's curse' bias, as results based on an external cohort not only did not improve on the original PGC-based effects, but they showed a very significant drop to the amount of variance explained by PRS.

How do my negative findings fit into the existing literature on this topic? Since

completing this work, in April 2023 an analysis from a Broad Institute group was published, asking exactly the same question, i.e., whether there was evidence of *nonadditive effects* genome-wide in more than 1,000 phenotypes in the UK Biobank population (Palmer et al., 2023). After analysing 361,194 samples for 13.7 million SNPs, and testing for associations with 1060 phenotypes, Palmer et al., 2023 found just 183 phenotype-locus pairs that were genome-wide significant at $p < 4.7 \times 10^{-11}$. They concluded that additive effects exist, but that they are rare, and that minimum sample sizes of millions are required to detect nonadditive effects at the same strength of association as those reported for additive effects. Palmer et al., 2023 did not test for schizophrenia, because they excluded binary traits with fewer than 3,000 cases (or controls), as this implied less than eight samples in both categories (case and controls) to be homozygous down to an allele frequency of 5% (which they had set as the minimum MAF for including a SNP in the analysis).

It is therefore evident that, despite having had the right instinct in performing this analysis, my main **limitation** was that I was not powered to detect dominant effects in the UK Biobank, which can count only on 1,169 people with schizophrenia (within the European ethnicity I had planned to analyse). Further, I found the largest effects in SNPs with MAFs < 0.05, which I should have instead excluded due to power constraints. Finally, Palmer et al., 2023 show that the nonadditive variance contributions decreases as the MAF decreases from 0.5, i.e., for rarer SNPs, even if a dominant effect is demonstrated (which – as discussed – requires very large samples sizes), this would not make significant contributions to the amount of variance explained by the model.

In **further work** I plan to repeat this analysis on dominant effects on the whole PGC schizophrenia cohort, to interrogate a much larger sample, powered to detect nonadditive effects. To do so, I will need to overcome the limitations posed by the PGC sample, the main of which is that it is composed of over 90 datasets, each with a maximum of a few thousands patients, therefore perpetuating some of the small-sample limitations that I have encountered in this work.

5.4 Conclusions

In this work I have tackled the issue of complex disorder 'missing heritability' through a diverse set of analyses. To test my hypotheses, I have used two model conditions, schizophrenia and hypertrophic cardiomyopathy (HCM). Schizophrenia was chosen as a highly polygenic disorder, while HCM appeared a good comparison because of lower polygenicity (it is a condition with a few very high-risk variants – including some for sarcomeric proteins – alongside a number of common, low risk variants). Further, while schizophrenia has been described as a multi-system, developmental disorder – with multiple insults converging on the synapse as the final common pathway (Howes & Onwordi, 2023; Pillinger et al., 2019a) – HCM is a more tissue-specific condition, whose genetic roots might fully exert their effects in the adult, fully differentiated tissue (Marian & Braunwald, 2017; Mazzarotto et al., 2020; Tadros et al., 2023).

In Chapter 2 I have found that, as hypothesised, neural-specific significant enhancer partitions – particularly when enhancers overlapped GRBs – were highly enriched in schizophrenia heritability. On the other hand, cardiac-specific significant enhancer partitions' enrichment in HCM heritability did not survive FDR correction.

In Chapter 3 I found that the use of multivariable logistic models including partitioned PRSs – particularly the model including NEURAL SIGNIFICANT ENHANCERS WITHIN GRBS and a *residual* partition as predictors, and including predictor interactions and quadratic terms – showed a 6.5% increase in the Coefficient of Determination for schizophrenia over the original leave-one-out GWAS. However, this increase did not appear statistically significant, due to the width of the confidence intervals, nor is probably clinically relevant for the production of improved risk prediction models for psychosis. However, the findings are in line with similar recent attempts from large groups, including for example Márquez-Luna et al., 2021.

For HCM, I found limited effects from the use of partitioned PRSs, and, especially in the UK Biobank sample, CoD increases following the use of quadratic terms – possibly due to the

fact that squaring predictors increased the weight of extreme PRS values, which appeared much more common in this less polygenic condition. This effect appeared similar but much smaller using the Royal Brompton cohort.
Lastly, in this chapter I have showed that, in a completely negative finding, and contrary to my expectations, enhancer-based annotations could not help improve PRSs.

In Chapter 4 I tested nonadditive effects as applied to enhancer-based SNPs, within an EP-WAS (an enhancer-based association study) of schizophrenia. For this analysis my results are inconclusive, as the sample I have used has been shown in the very recent literature to be insufficiently powered to detect nonadditive effects (Palmer et al., 2023). My results also highlight one of the reasons additive scores are normally considered in most GWASes: they require smaller sample sizes to be powered to detect associations. I plan on repeating this analysis in larger samples as part of my future research endeavours, once I can resolve a few important difficulties based on the PGC's sample structure.

Overall, pending further validation for some of the findings, my research points to the fact that:

1. Selecting neural tissue-expressed enhancers does select for genomic areas of increased importance for schizophrenia heritability, especially when these overlap GRBs.
2. Partitioned PRSs could potentially modestly increase schizophrenia's coefficient of determination – however this does not seem to be an approach that can revolutionise the field.
3. When modelling the risk for less polygenic conditions such as HCM, it might be worth considering both linear and quadratic terms.
4. Nonadditive effect should be explored in GWAS – however this requires very large samples.

Appendix A

Supplementary Figures

A.1 Chapter 3 – Sensitivity analysis: Neural tissue and schizophrenia – additional *clz2a* and *celso* cohorts at 0.5 threshold

A.1.1 *clz2a* cohort

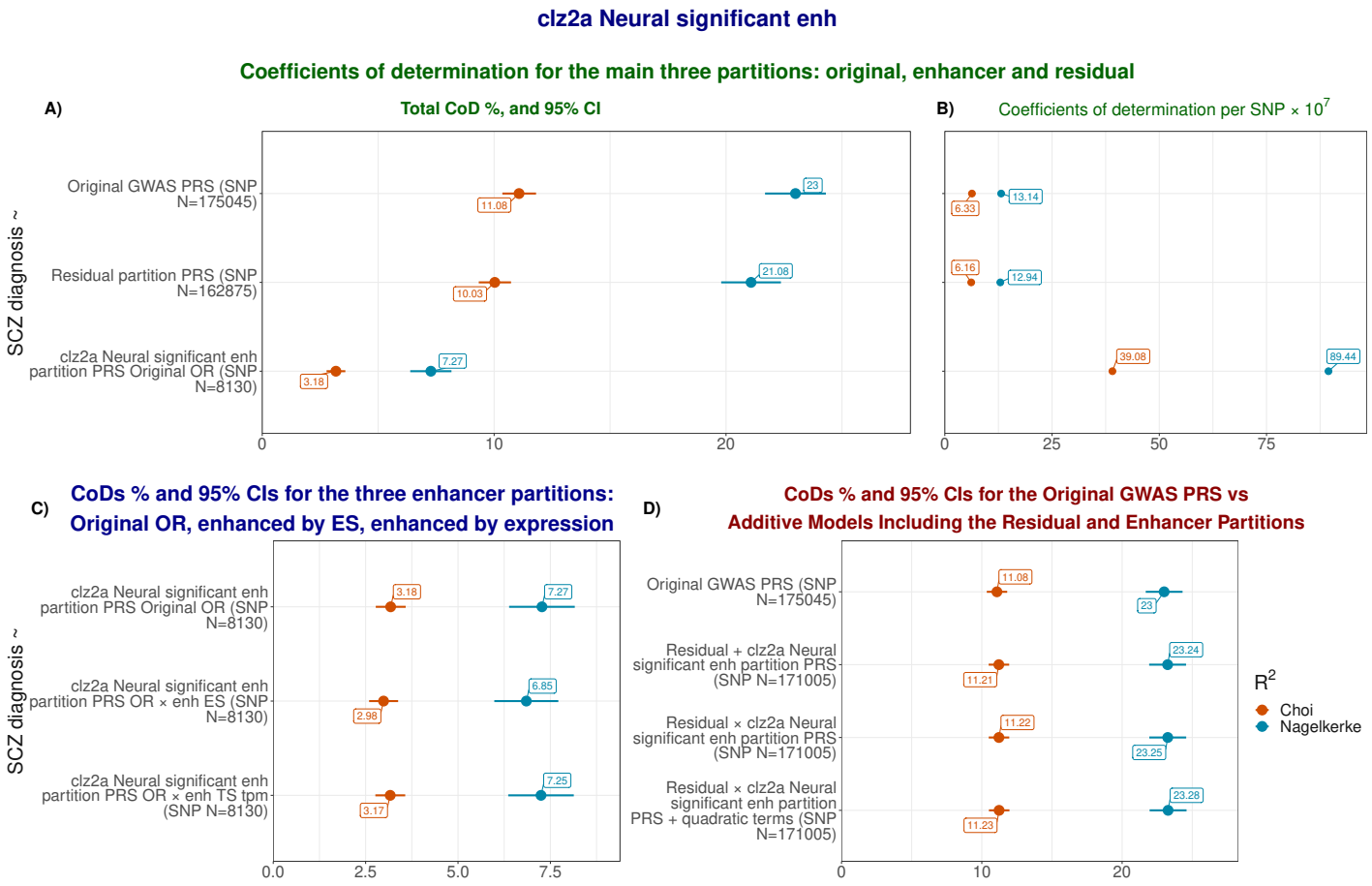


Figure A.1: Coefficients of Determination for Schizophrenia for Neural significant enhancers, *clz2a* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models. In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the clz2a cohort. Target data: clz2a European PGC schizophrenia cohort.

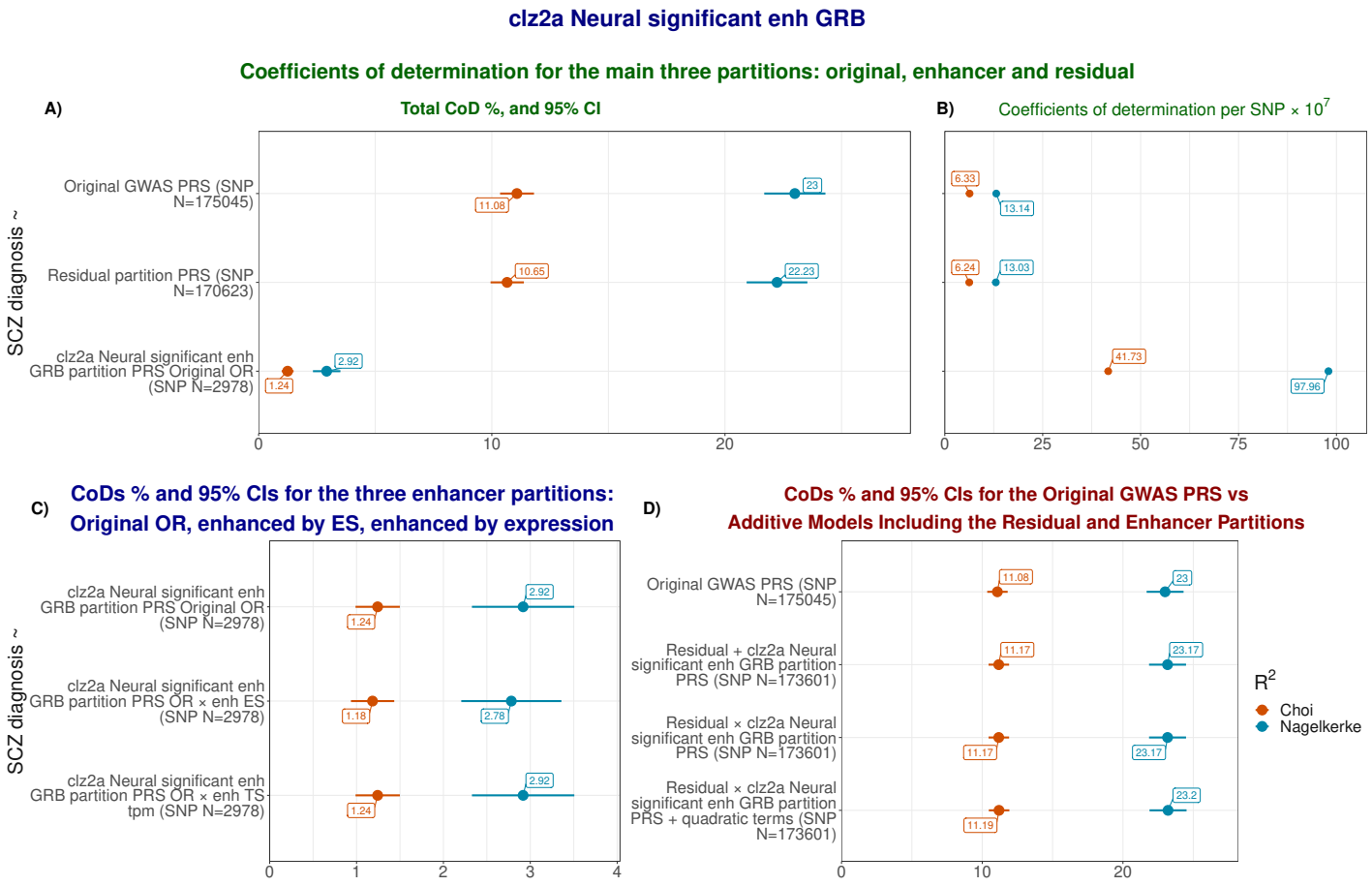


Figure A.2: Coefficients of Determination for Schizophrenia for Neural significant enhancers within GRBs, *clz2a* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the clz2a cohort. Target data: clz2a European PGC schizophrenia cohort.

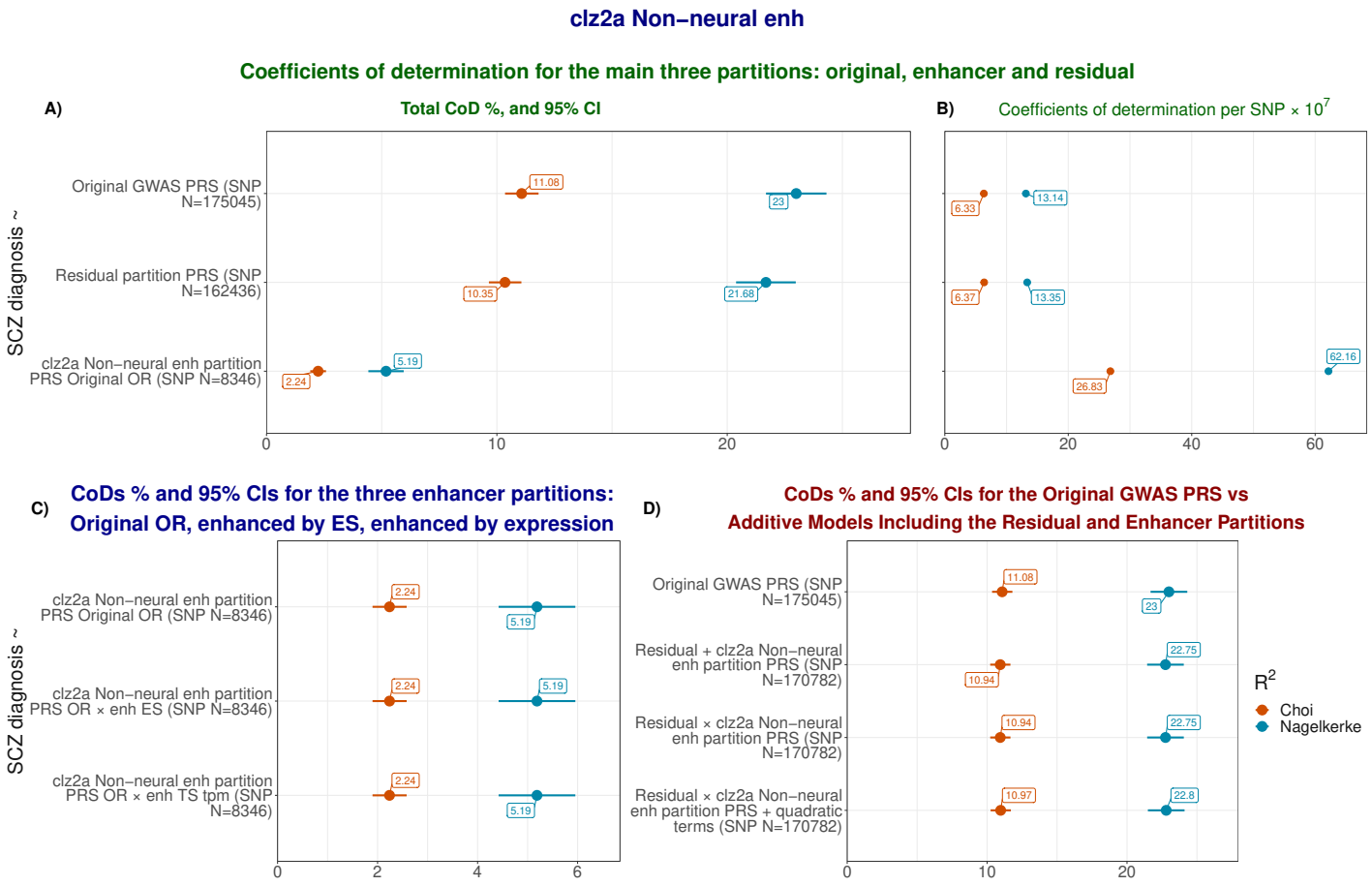


Figure A.3: Coefficients of Determination for Schizophrenia for Non neural enhancers, *clz2a* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the clz2a cohort. Target data: clz2a European PGC schizophrenia cohort.

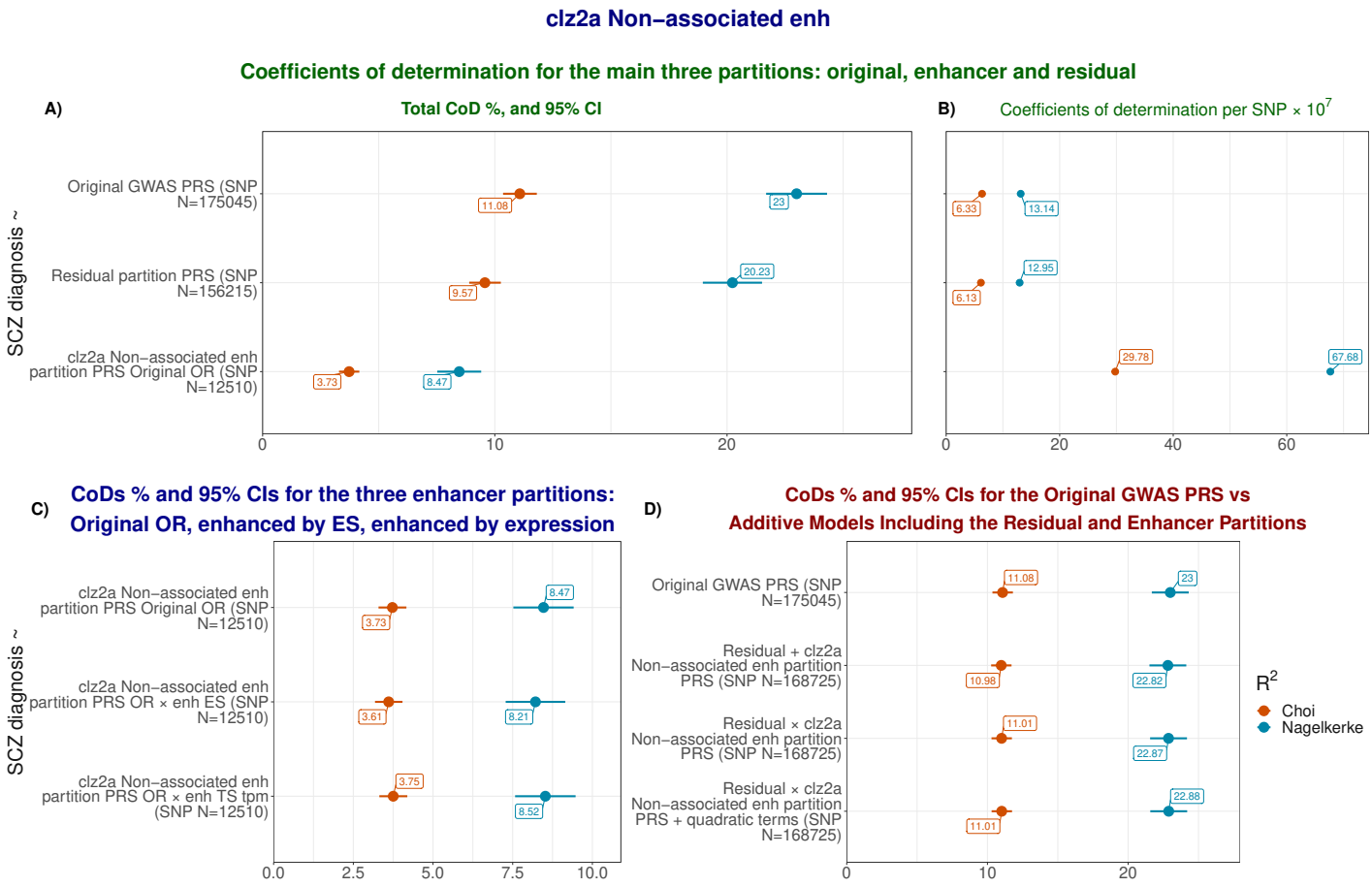


Figure A.4: Coefficients of Determination for Schizophrenia for Non associated enhancers, *clz2a* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the clz2a cohort. Target data: clz2a European PGC schizophrenia cohort.

A.1.2 *celso* cohort

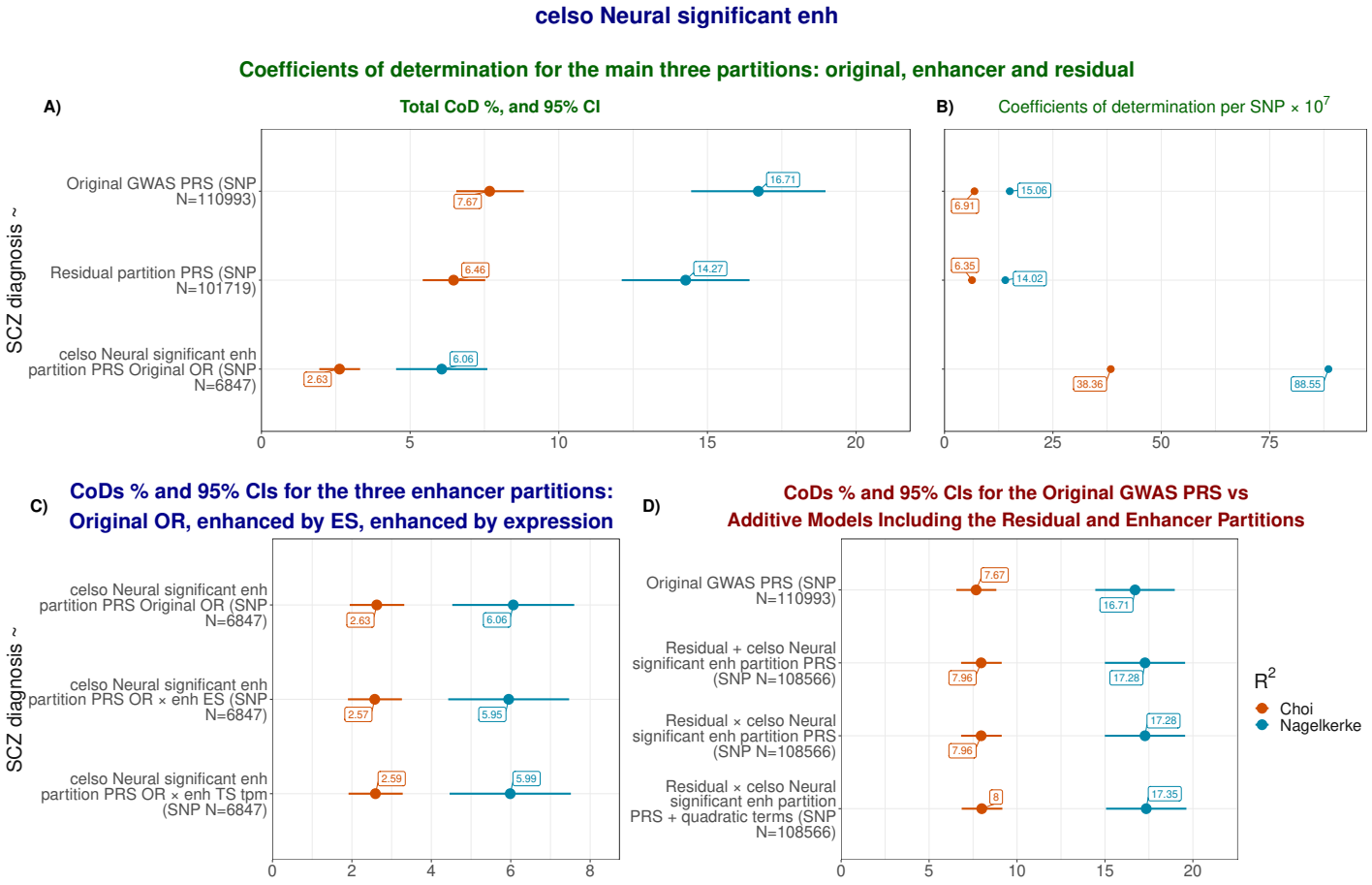


Figure A.5: Coefficients of Determination for Schizophrenia for Neural significant enhancers, *celso* cohort.
Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.
In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.
Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the celso cohort. Target data: celso European PGC schizophrenia cohort.

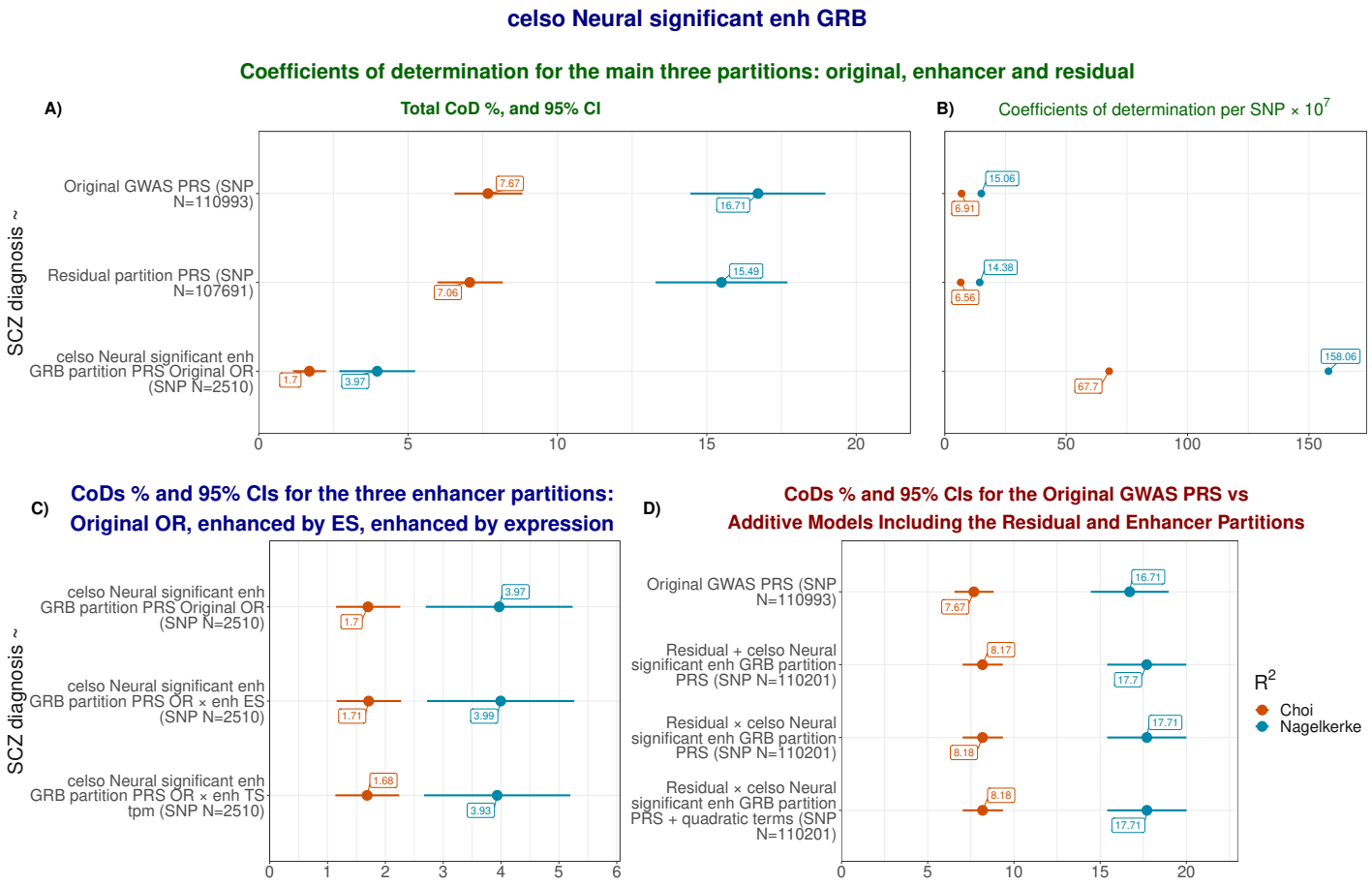


Figure A.6: Coefficients of Determination for Schizophrenia for Neural significant enhancers within GRBs, *celso* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the celso cohort. Target data: celso European PGC schizophrenia cohort.

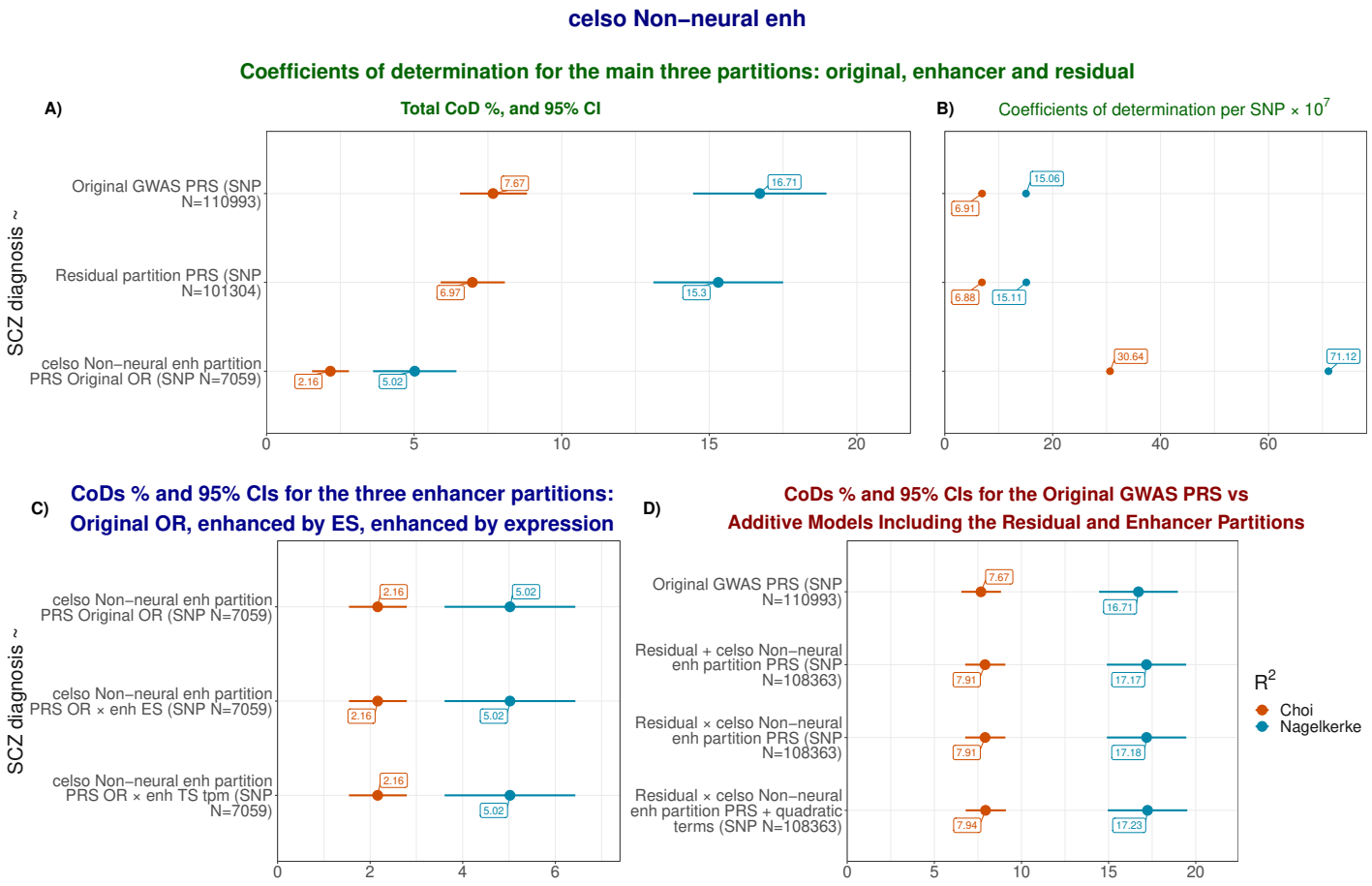


Figure A.7: Coefficients of Determination for Schizophrenia for Non neural enhancers, *celso* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the celso cohort. Target data: celso European PGC schizophrenia cohort.

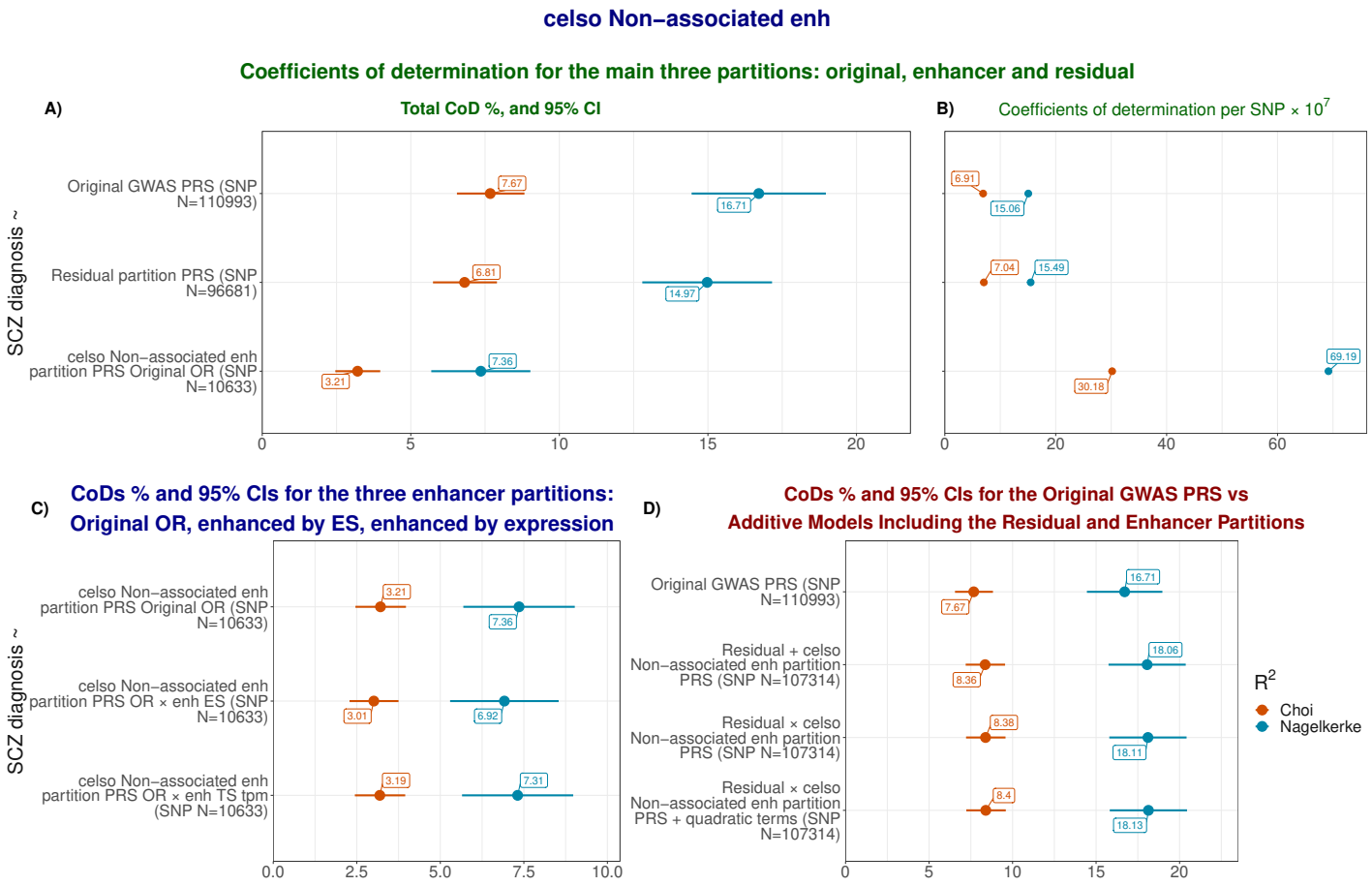


Figure A.8: Coefficients of Determination for Schizophrenia for Non associated enhancers, *celso* cohort.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the celso cohort. Target data: celso European PGC schizophrenia cohort.

A.2 Chapter 3 – Sensitivity analysis: Neural tissue and schizophrenia – *xs234* cohort, 0.05 *p*-value threshold

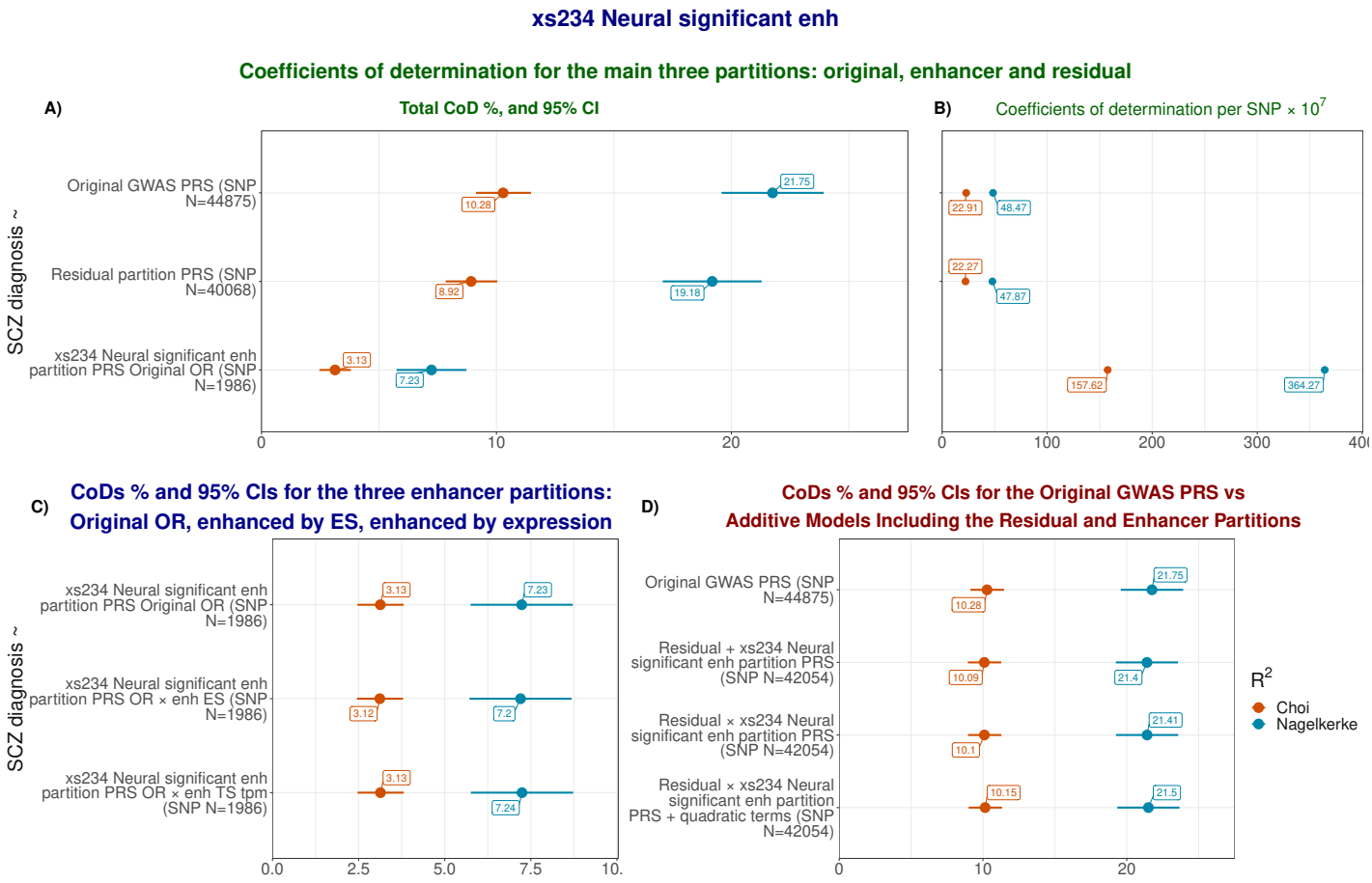


Figure A.9: Coefficients of Determination for Schizophrenia for Neural significant enhancers, *xs234* cohort, 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

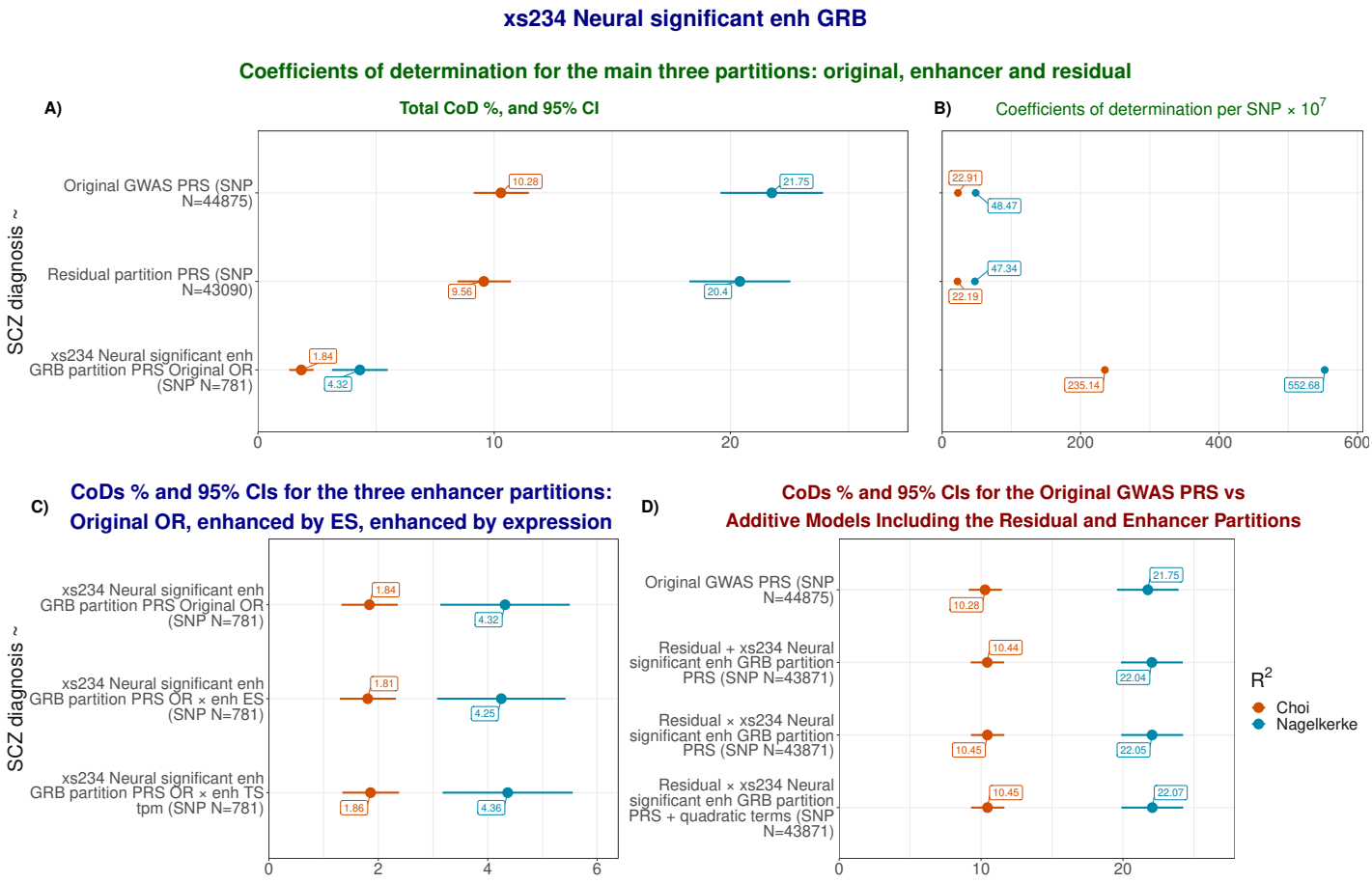


Figure A.10: Coefficients of Determination for Schizophrenia for Neural significant enhancers within GRBs, *xs234* cohort, 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

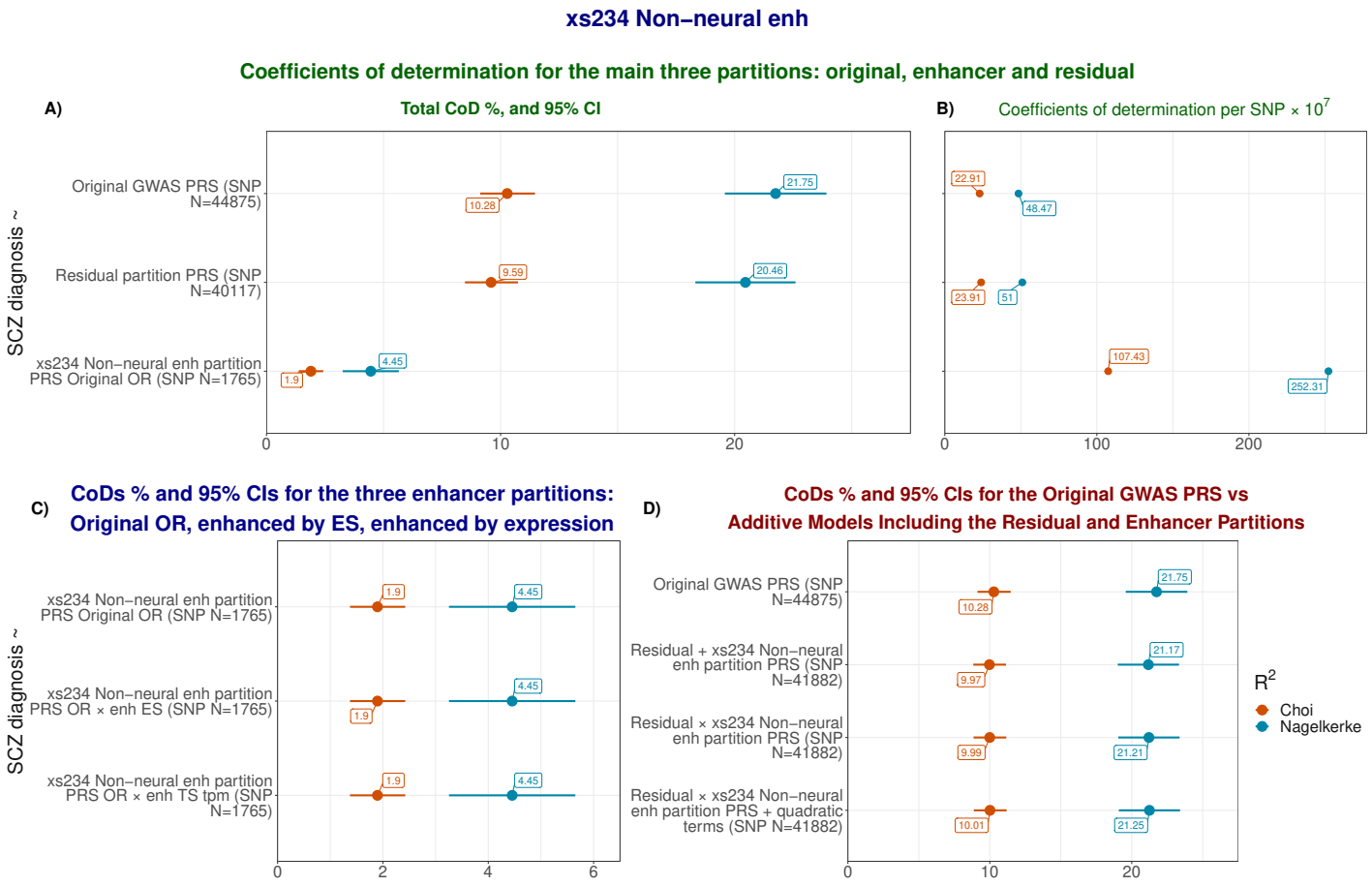


Figure A.11: Coefficients of Determination for Schizophrenia for Non neural enhancers, *xs234* cohort, 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

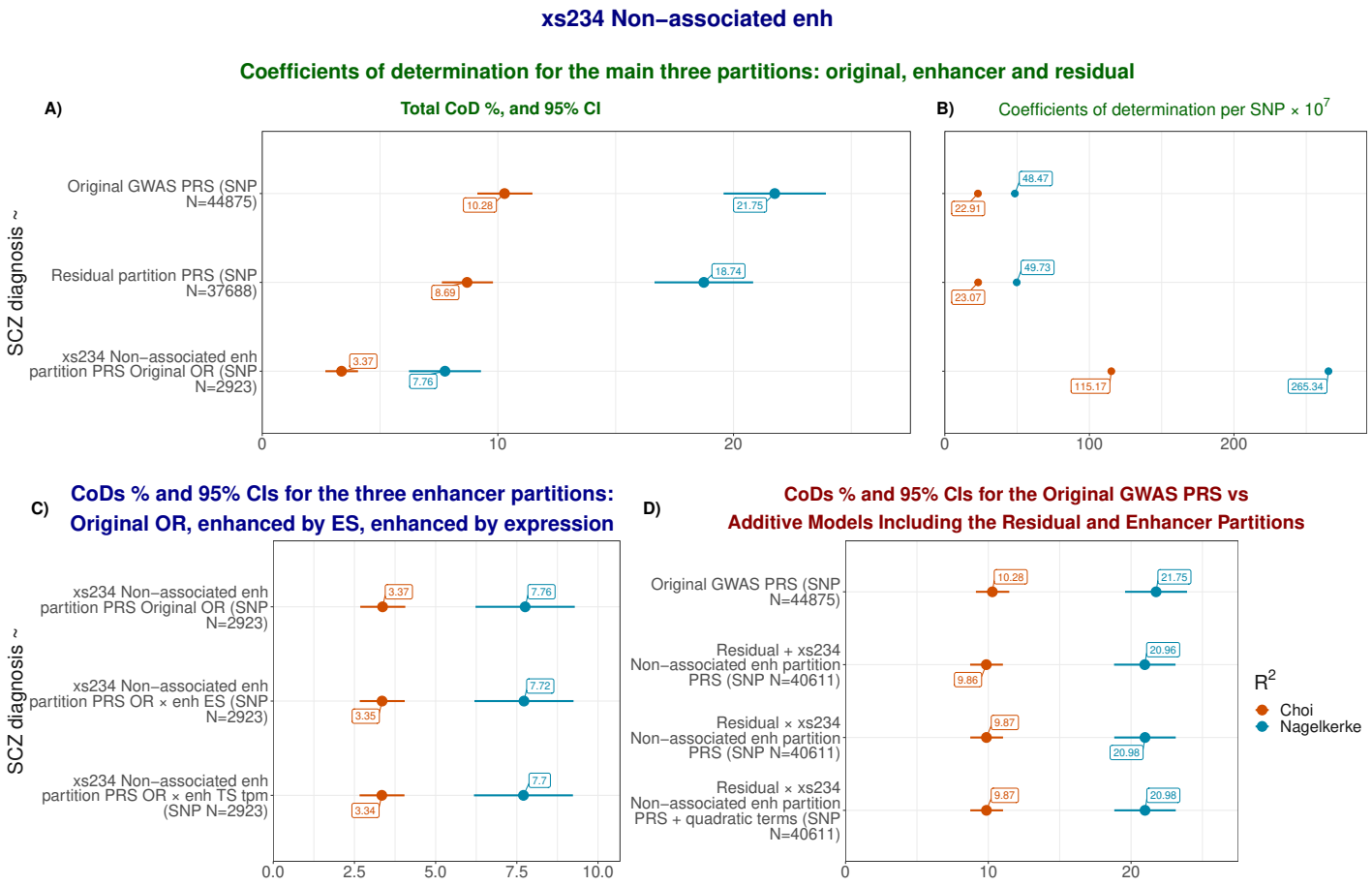


Figure A.12: Coefficients of Determination for Schizophrenia for Non associated enhancers, *xs234* cohort, 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original LOO GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: leave-one-out (LOO) European ancestry PGC GWAS for schizophrenia for the xs234 cohort. Target data: xs234 European PGC schizophrenia cohort.

A.3 Chapter 3 – Sensitivity analysis: Cardiac tissue and HCM – 0.05 threshold

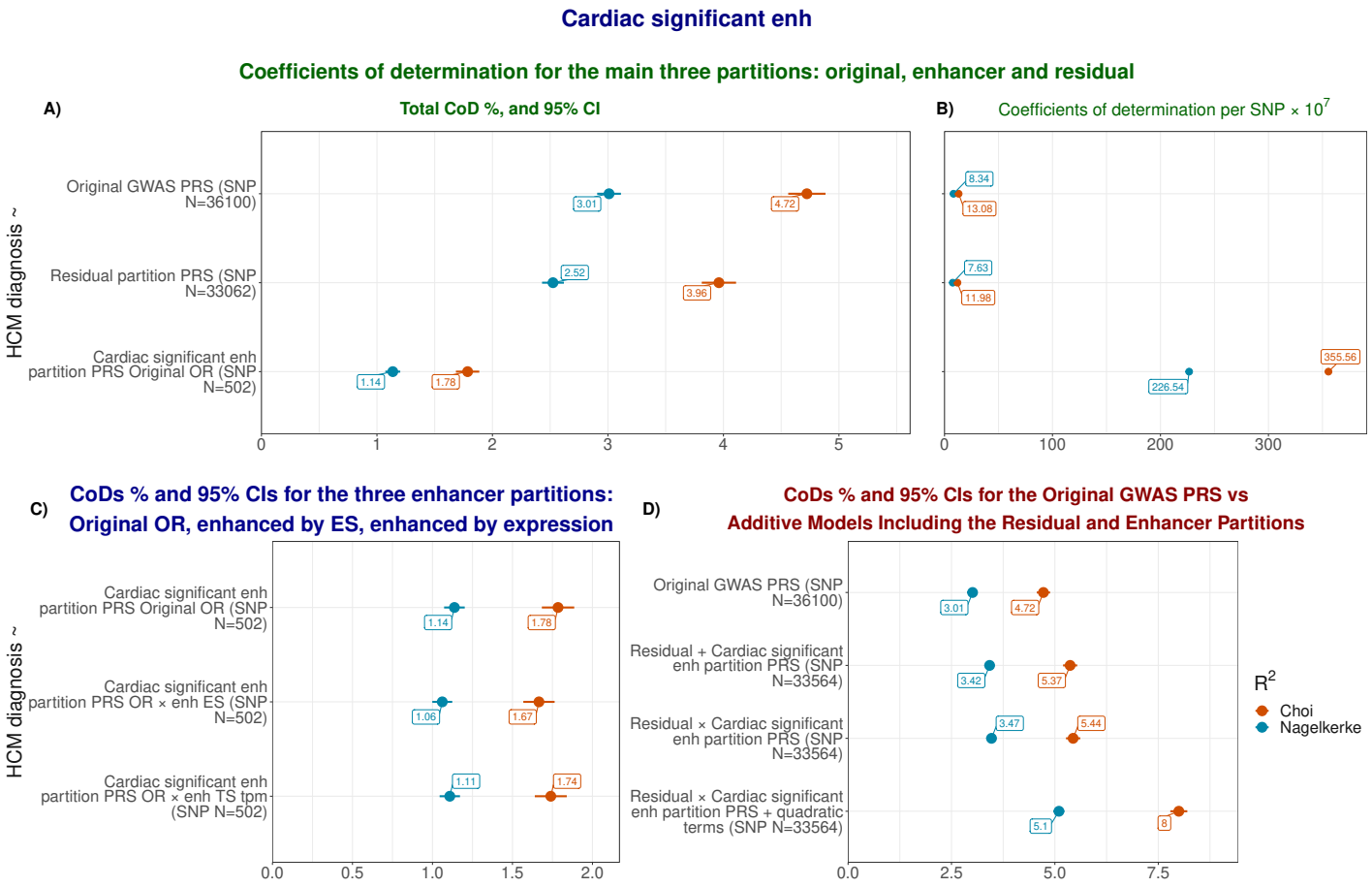


Figure A.13: Coefficients of determination for HCM for Cardiac significant enhancers, in the *UKBB* cohort – 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

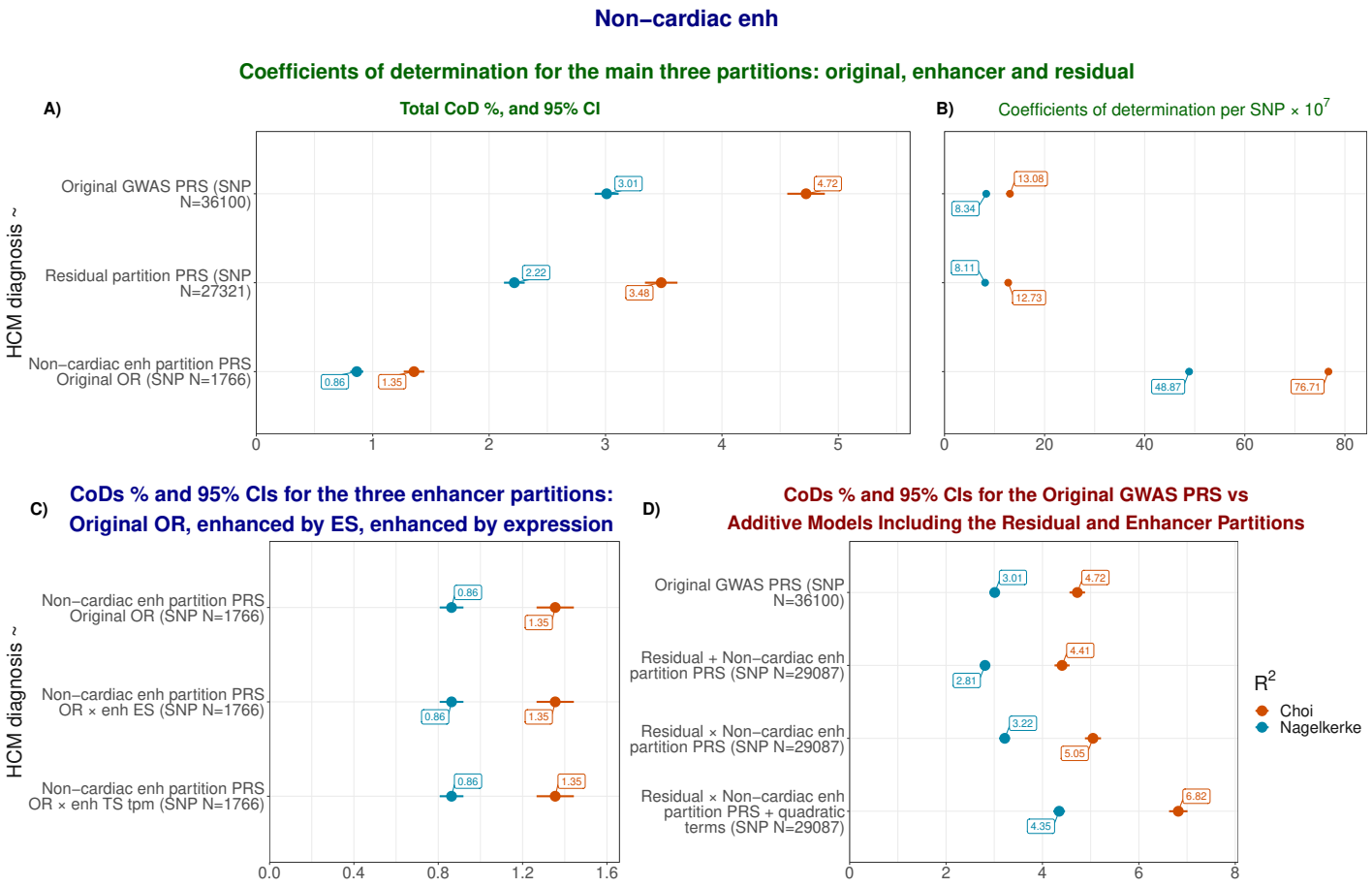


Figure A.14: Coefficients of determination for HCM for Non cardiac enhancers, in the *UKBB* cohort – 0.05 threshold.

Panel A) shows the coefficients of determination and 95% confidence intervals for the three main partitions (original GWAS, residual, and enhancer-based) in this cohort. **Panel B)** shows the corresponding point values, adjusted per SNP ($\times 10^7$). **Panel C)** shows the coefficients of determination and 95% confidence intervals for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer. **Panel D)** shows the coefficients of determination and 95% confidence intervals for each PRS for the original GWAS PRS, as well as for the three partitions using additive models.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original *Nagelkerke's* R^2 for comparison.

Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

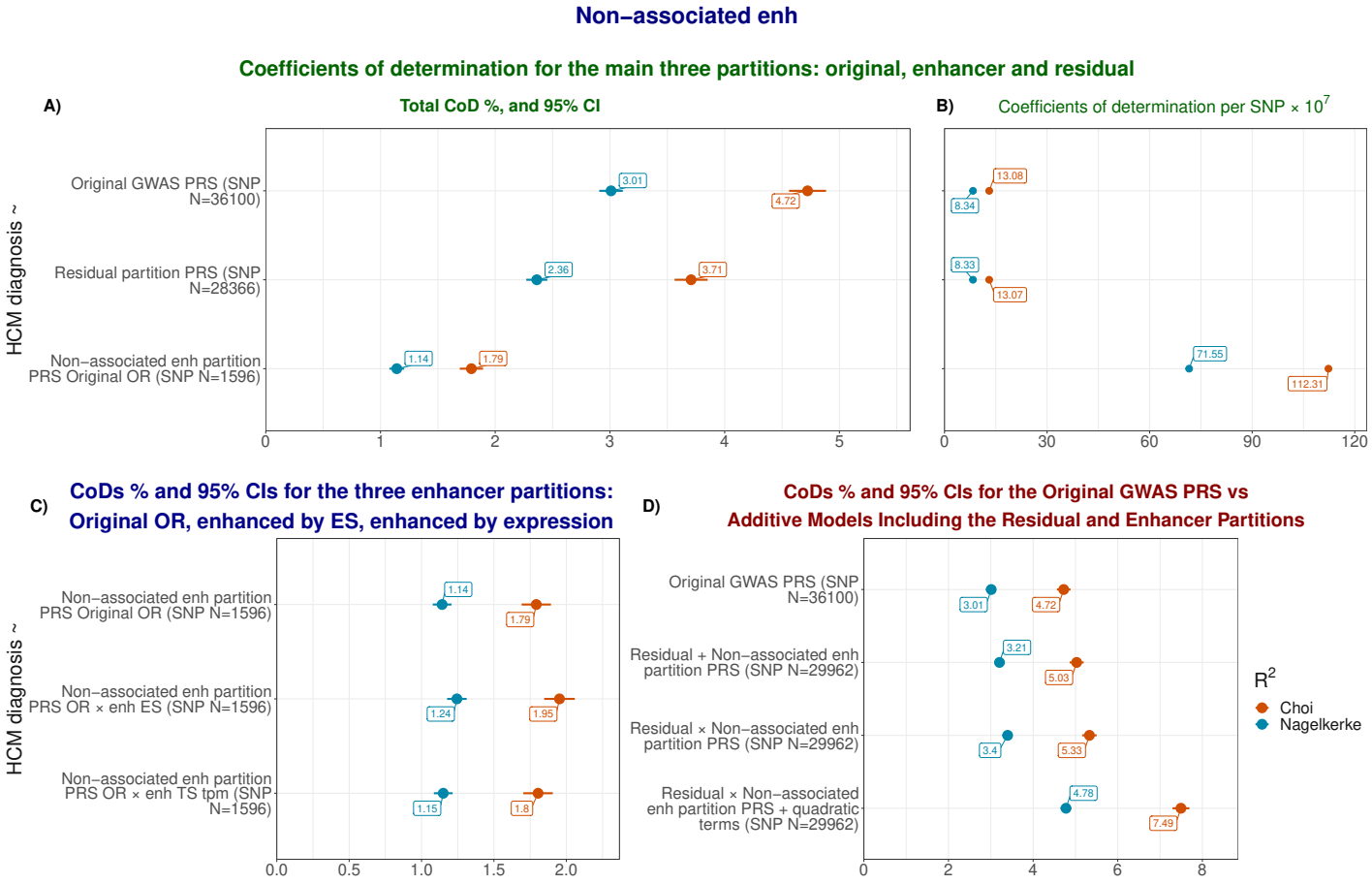


Figure A.15: Coefficients of determination for HCM for Non associated enhancers, in the *UKBB* cohort – 0.05 threshold.

The figure describes the proportion of the variance of HCM explained by the genetic factor for each PRS for the three enhancer-based partitions – one based on the original OR, the second based on the OR enhanced by the effect size (ES) of association between each enhancer and promoter, and the final one based on the OR enhanced by the tissue-specific expression value of each enhancer.

In brick red, values are on the liability scale, corrected for ascertainment as per Lee et al., 2012, utilising the formula by Choi and O'Reilly, 2019 – or coefficients of determination (CoD). In baby blue the original Nagelkerke's R^2 for comparison. Plots represent the CoD % and 95% confidence interval for the measure. **Panel A)** shows the CoD for each genomic partition for the CARDIAC SIGNIFICANT list. **Panel B)** shows the CoD for each genomic partition for the NON CARDIAC list. **Panel C)** shows the CoD for each genomic partition for the NON ASSOCIATED list.

Base data: HCM GWAS by Tadros et al., 2023. Target data: UKBB European sample.

References

Abdellaoui, A., & Verweij, K. (2022). Genetica en psychiatrie. *Tijdschrift voor Psychiatrie*, (2022/5), 260–265 (cit. on p. 33).

Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J. C., Suzuki, H., Daub, C. O., Hayashizaki, Y., & Lenhard, B. (2009). Transcriptional features of genomic regulatory blocks. *Genome biology*, *10*(4), R38 (cit. on pp. 14, 44, 73).

Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., Jaffe, A. E., Pinto, D., Dracheva, S., & Geschwind, D. H. (2015). The PsychENCODE project. *Nature neuroscience*, *18*(12), 1707–1712 (cit. on pp. 51, 140).

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics*, *11*(8), 559–571 (cit. on pp. 1, 2).

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K. K., Matsuda, K., Murakami, Y., Price, A. L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nature genetics*, *52*(12), 1346–1354 (cit. on p. 144).

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., & Suzuki, T. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461 (cit. on pp. 2, 9, 16, 53, 56, 61, 63, 64, 67, 68, 137, 139, 146).

Bagos, P. G. (2013). Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis. *Statistical Applications in Genetics and Molecular Biology*, *12*(3), 285–308 (cit. on pp. 23, 24, 119).

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature reviews genetics*, *7*(10), 781–791 (cit. on pp. 22, 23, 119).

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–740 (cit. on p. 8).

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308 (cit. on p. 8).

Barešić, A., Nash, A. J., Dahoun, T., Howes, O., & Lenhard, B. (2020). Understanding the genetics of neuropsychiatric disorders: The potential role of genomic regulatory blocks. *Molecular Psychiatry*, *25*(1), 6–18 (cit. on pp. 16, 48, 50, 73, 135, 137–139, 142).

Battle, A., Brown, C. D., Engelhardt, B. E., & Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213 (cit. on p. 53).

Beagan, J. A., & Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nature genetics*, *52*(1), 8–16 (cit. on p. 6).

Becker, T. S., & Rinkwitz, S. (2012). Zebrafish as a genomics model for human neurological and polygenic disorders. *Developmental neurobiology*, *72*(3), 415–428 (cit. on p. 12).

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, *304*(5675), 1321–1325 (cit. on pp. 11, 12).

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276 (cit. on p. 7).

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678 (cit. on pp. 5, 6).

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, *144*(3), 327–339 (cit. on p. 9).

Bulik-Sullivan, B., Finucane, H. K., Neale, B., & Price, A. L. (2015). LD score regression (LDSC). (Cit. on pp. 48, 49, 55, 56).

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., & O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209 (cit. on pp. 77, 102, 120).

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, *309*(5740), 1559–1563 (cit. on p. 2).

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015-0047–8 (cit. on p. 85).

Chatterjee, N., Shi, J., & Garcia-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, *17*(7), 392–406 (cit. on pp. 71, 72).

Chen, J., & Chatterjee, N. (2007). Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. *Human heredity*, *63*(3-4), 196–204 (cit. on p. 23).

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: A guide to performing polygenic risk score analyses. *Nature protocols*, *15*(9), 2759–2772 (cit. on pp. 28, 29).

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic risk score software for biobank-scale data. *GigaScience*, *8*(7) (cit. on pp. 30, 82, 84, 85, 89, 90, 93, 94, 98, 104, 106, 109, 110, 112, 113, 127, 128, 131, 132, 152–159, 161–164, 166–168).

Chua, E. H. Z., Yasar, S., & Harmston, N. (2022). The importance of considering regulatory domains in genome-wide analyses–the nearest gene is often wrong! *Biology Open*, *11*(4), bio059091 (cit. on p. 14).

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., & Shumway, M. (2012). The 1000 genomes project: Data management and community access. *Nature methods*, *9*(5), 459–462 (cit. on p. 55).

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., et al. (2020). A brief history of human disease genetics. *Nature*, *577*(7789), 179–189 (cit. on pp. 27, 28).

Cloutier, M., Aigbogun, M. S., Guerin, A., Nitulescu, R., Ramanakumar, A. V., Kamat, S. A., DeLucia, M., Duffy, R., Legacy, S. N., Henderson, C., et al. (2016). The economic burden of schizophrenia in the United States in 2013. *The Journal of clinical psychiatry*, *77*(6), 5379 (cit. on p. 35).

Coelewij, L., & Curtis, D. (2018). Mini-review: Update on the genetics of schizophrenia. *Annals of human genetics*, *82*(5), 239–243 (cit. on pp. 38, 40).

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The human genome project: Lessons from large-scale biology. *Science*, *300*(5617), 286–290 (cit. on p. 2).

Dawn Teare, M., & Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, *366*(9490), 1036–1044 (cit. on p. 21).

De La Calle-Mustienes, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodriguez-Seguel, E., Letizia, A., Allende, M. L., & Gómez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate iroquois cluster gene deserts. *Genome research*, *15*(8), 1061–1072 (cit. on p. 11).

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., & Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology*, *8*(5), e1000384 (cit. on p. 9).

De Wit, E., & De Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes & development*, *26*(1), 11–24 (cit. on p. 7).

Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'shea, C. C., Park, P. J., & Ren, B. (2017). The 4D nucleome project. *Nature*, *549*(7671), 219–226 (cit. on p. 51).

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, *295*(5558), 1306–1311 (cit. on p. 7).

de Villiers, J., & Schaffner, W. (1981). A small segment of polyoma virus DNA enhances the expression of a cloned β-globin gene over a distance of 1400 base pairs. *Nucleic acids research*, *9*(23), 6251–6264 (cit. on p. 8).

DeWan, A., Liu, M., Hartman, S., Zhang, S. S.-M., Liu, D. T., Zhao, C., Tam, P. O., Chan, W. M., Lam, D. S., Snyder, M., et al. (2006). HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science*, *314*(5801), 989–992 (cit. on p. 134).

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, *35*(4), 316–319 (cit. on p. 85).

Dimitrieva, S., & Bucher, P. (2013). UCNEbase – a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research*, *41*(D1), D101–D109 (cit. on p. 11).

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380 (cit. on p. 4).

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108 (cit. on pp. 2, 9).

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, *9*(3), e1003348 (cit. on p. 28).

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomaRt. *Nature protocols*, *4*(8), 1184–1191 (cit. on p. 85).

Elliott, J., Bodinier, B., Bond, T. A., Chadeau-Hyam, M., Evangelou, E., Moons, K. G., Dehghan, A., Muller, D. C., Elliott, P., & Tzoulaki, I. (2020). Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA*, *323*(7), 636–645 (cit. on p. 71).

Engström, P. G., Fredman, D., & Lenhard, B. (2008). Ancora: A web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome biology*, *9*, 1–12 (cit. on pp. 11, 12).

Engström, P. G., Sui, S. J. H., Drivenes, Ø., Becker, T. S., & Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome research*, *17*(12), 1898–1908 (cit. on p. 11).

Esteller, M. (2011). Non-coding rnas in human disease. *Nature reviews genetics*, *12*(12), 861–874 (cit. on p. 2).

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics*, *31*(9), 1466–1468 (cit. on pp. 28–30, 82).

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., & Farh, K. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, *47*(11), 1228–1235 (cit. on pp. 49, 55, 56, 58).

Flavahan, W. A., Drier, Y., Johnstone, S. E., Hemming, M. L., Tarjan, D. R., Hegazi, E., Shareef, S. J., Javed, N. M., Raut, C. P., Eschle, B. K., et al. (2019). Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature*, *575*(7781), 229–233 (cit. on p. 14).

Flöttmann, R., Wagner, J., Kobus, K., Curry, C. J., Savarirayan, R., Nishimura, G., Yasui, N., Spranger, J., Van Esch, H., & Lyons, M. J. (2015). Microdeletions on 6p22. 3 are associated with mesomelic dysplasia savarirayan type. *Journal of medical genetics*, *52*(7), 476–483 (cit. on p. 6).

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., & Patwardhan, T. A. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature genetics*, *51*(12), 1664–1669 (cit. on p. 16).

Gebert, L. F., & MacRae, I. J. (2019). Regulation of microRNA function in animals. *Nature reviews Molecular cell biology*, *20*(1), 21–37 (cit. on p. 2).

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., & Gentry, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), 1–16 (cit. on p. 85).

Georgieva, R. (2022). *Responsiveness of genes to long-range transcriptional regulation* (Thesis). (Cit. on pp. 13–18, 48–50, 52–54, 56–58, 62, 73, 78, 137–140, 142, 146).

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., & Atzori, C. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: A second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human molecular genetics*, *24*(11), 3143–3154 (cit. on p. 6).

Gong, G., Hannon, N., & Whittemore, A. S. (2010). Estimating gene penetrance from family data. *Genetic epidemiology*, *34*(4), 373–381 (cit. on p. 22).

Grant, S. W., Collins, G. S., & Nashef, S. A. (2018). Statistical primer: Developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery*, *54*(2), 203–208 (cit. on p. 70).

Haberle, V., Forrest, A. R., Hayashizaki, Y., Carninci, P., & Lenhard, B. (2015). CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic acids research*, *43*(8), e51–e51 (cit. on p. 56).

Haberle, V., & Lenhard, B. (2016). Promoter architectures and developmental gene regulation. *Seminars in cell & developmental biology*, *57*, 11–23 (cit. on p. 8).

Hamada, M., Ikeda, S., & Shigematsu, Y. (2014). Advances in medical treatment of hypertrophic cardiomyopathy. *Journal of cardiology*, *64*(1), 1–10 (cit. on p. 41).

Harmston, N., Barešić, A., & Lenhard, B. (2013). The mystery of extreme non-coding conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1632), 20130021 (cit. on p. 11).

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., & Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with metazoan clusters of extreme noncoding conservation. *Nature communications*, *8*(1), 1–13 (cit. on pp. 12, 52, 54, 62).

Harmston, N., & Lenhard, B. (2013). Chromatin and epigenetic features of long-range gene regulation. *Nucleic acids research*, *41*(15), 7185–7199 (cit. on p. 4).

Harrison, P. J. (2015). Recent genetic findings in schizophrenia and their therapeutic relevance. *Journal of psychopharmacology*, *29*(2), 85–96 (cit. on p. 40).

Hatzis, P., & Talianidis, I. (2002). Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Molecular cell*, *10*(6), 1467–1477 (cit. on p. 10).

Heun, P., Laroche, T., Shimada, K., Furrer, P., & Gasser, S. M. (2001). Chromosome dynamics in the yeast interphase nucleus. *Science*, *294*(5549), 2181–2186 (cit. on p. 7).

Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T. M., Nordentoft, M., & Glenthøj, B. (2018). Heritability of schizophrenia and schizophrenia spectrum based on the nationwide danish twin register. *Biological psychiatry*, *83*(6), 492–498 (cit. on p. 32).

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: Prospective open cohort study. *BMJ*, *335*(7611), 136 (cit. on p. 71).

Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: A systematic review and meta-analysis. *The Lancet Psychiatry*, *4*(4), 295–301 (cit. on p. 35).

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., & Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934–947 (cit. on pp. 61, 63, 67).

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., & Birney, E. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, *41*(2), 827–841 (cit. on pp. 61, 63, 67).

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III – the final common pathway. *Schizophrenia bulletin*, *35*(3), 549–562 (cit. on pp. 37, 38).

Howes, O. D., & Onwordi, E. C. (2023). The synaptic hypothesis of schizophrenia version III: A master mechanism. *Molecular Psychiatry*, 1–14 (cit. on pp. 37, 48, 74, 138, 149).

Hraba-Renevey, S., & Kress, M. (1989). Expression of a mouse replacement histone H3.3 gene with a highly conserved 3' noncoding region during SV40- and polyoma-induced Go to S-phase transition. *Nucleic acids research*, *17*(7), 2449–2461 (cit. on p. 11).

Hsieh, T.-H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell*, *162*(1), 108–119 (cit. on p. 7).

Jannot, A.-S., Ehret, G., & Perneger, T. (2015). P 5×10^{-8} has emerged as a standard of statistical significance for genome-wide association studies. *Journal of clinical epidemiology*, *68*(4), 460–465 (cit. on p. 24).

Jostins, L., & Barrett, J. C. (2011). Genetic risk prediction in complex disease. *Human molecular genetics*, *20*(R2), R182–R188 (cit. on p. 21).

Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., & Chee, S. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature genetics*, *51*(10), 1442–1449 (cit. on p. 16).

Kalpana, B., Murthy, D. K., Balakrishna, N., & Aiyengar, M. T. (2019). Genetic variants of chromosome 9p21. 3 region associated with coronary artery disease and premature coronary artery disease in an Asian Indian population. *Indian Heart Journal*, *71*(3), 263–271 (cit. on p. 26).

Karayiorgou, M., Morris, M. A., Morrow, B., Shprintzen, R. J., Goldberg, R., Borrow, J., Gos, A., Nestadt, G., Wolyniec, P. S., & Lasseter, V. K. (1995). Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proceedings of the National Academy of Sciences*, *92*(17), 7612–7616 (cit. on p. 39).

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., & Ellinor, P. T. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, *50*(9), 1219–1224 (cit. on p. 31).

Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B., & Becker, T. S. (2007a). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks-a fundamental feature of vertebrate genomes. *Genome biology*, *8*, 1–13 (cit. on pp. 11, 12).

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., & Howe, K. (2007b). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*, *17*(5), 545–555 (cit. on pp. 11, 12).

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187 (cit. on p. 9).

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389 (cit. on p. 134).

Kleinjan, D. A., & Van Heyningen, V. (2005). Long-range control of gene expression: Emerging mechanisms and disruption in disease. *The American Journal of Human Genetics*, *76*(1), 8–32 (cit. on p. 12).

Krietenstein, N., Abraham, S., Venev, S. V., Abdennur, N., Gibcus, J., Hsieh, T.-H. S., Parsi, K. M., Yang, L., Maehr, R., & Mirny, L. A. (2020). Ultrastructural details of mammalian chromosome architecture. *Molecular cell*, *78*(3), 554–565. e7 (cit. on p. 51).

Langefeld, C. D., & Fingerlin, T. E. (2007). Association methods in human genetics. *Topics in Biostatistics*, 431–460 (cit. on p. 23).

Lawrence, M., Gentleman, R., & Carey, V. (2009). Rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, *25*(14), 1841–1842 (cit. on p. 85).

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS computational biology*, *9*(8), e1003118 (cit. on p. 85).

Lee, B. H., Wu, Z., & Rhie, S. K. (2022). Characterizing chromatin interactions of regulatory elements and nucleosome positions, using Hi-C, Micro-C, and promoter capture Micro-C. *Epigenetics and Chromatin*, *15*(1), 41 (cit. on p. 7).

Lee, S. H., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic epidemiology*, *36*(3), 214–224 (cit. on pp. 83, 89, 90, 93, 94, 98, 104, 106, 109, 110, 112, 113, 127, 128, 131, 132, 152–159, 161–164, 166–168).

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, *13*(4), 233–245 (cit. on p. 8).

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., & de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, *12*(14), 1725–1735 (cit. on p. 12).

Lewontin, R. C. (1964). The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, *49*(1), 49 (cit. on p. 26).

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in swedish families: A population-based study. *The Lancet*, *373*(9659), 234–239 (cit. on pp. 32, 135).

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England journal of medicine*, *353*(12), 1209–1223 (cit. on p. 38).

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., & Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293 (cit. on pp. 4, 7).

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, *389*(6648), 251–260 (cit. on p. 4).

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., & Laxova, R. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, *161*(5), 1012–1025 (cit. on pp. 6, 14).

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, *456*(7218), 18–21 (cit. on p. 32).

Mangalore, R., & Knapp, M. (2007). Cost of schizophrenia in England. *The journal of mental health policy and economics*, *10*(1), 23–41 (cit. on p. 35).

Manna, F., Martin, G., & Lenormand, T. (2011). Fitness landscapes: An alternative theory for the dominance of mutation. *Genetics*, *189*(3), 923–937 (cit. on pp. 46, 119, 147).

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, *27*(2), e1608 (cit. on p. 27).

Marian, A. J., & Braunwald, E. (2017). Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circulation research*, *121*(7), 749–770 (cit. on pp. 82, 149).

Maron, B. J., & Maron, M. S. (2013). Hypertrophic cardiomyopathy. *The Lancet*, *381*(9862), 242–255 (cit. on p. 41).

Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., & Price, A. L. (2021). Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications*, *12*(1), 6052 (cit. on pp. 144, 146, 149).

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature genetics*, *49*(1), 27–35 (cit. on p. 39).

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195 (cit. on pp. 4, 14).

Mazzarotto, F., Olivotto, I., Boschi, B., Girolami, F., Poggesi, C., Barton, P. J., & Walsh, R. (2020). Contemporary insights into the genetics of hypertrophic cardiomyopathy: Toward a new era in clinical testing? *Journal of the American Heart Association*, *9*(8), e015473 (cit. on pp. 41, 135, 139, 149).

McCutcheon, R. A., Abi-Dargham, A., & Howes, O. D. (2019). Schizophrenia, dopamine and the striatum: From biology to symptoms. *Trends in neurosciences*, *42*(3), 205–220 (cit. on pp. 37, 138).

McCutcheon, R. A., Marques, T. R., & Howes, O. D. (2020). Schizophrenia – an overview. *JAMA psychiatry*, *77*(2), 201–210 (cit. on pp. 35, 36, 138).

McVean, G., Spencer, C. C. A., & Chaix, R. (2005). Perspectives on human genetic variation from the HapMap project. *PLoS Genetics*, *1*(4), e54 (cit. on p. 27).

Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B. M., Rolando, D. M., Farabella, I., Morgan, C. C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature genetics*, *51*(7), 1137–1148 (cit. on p. 12).

Millar, J. K., Wilson-Annan, J. C., Anderson, S., Christie, S., Taylor, M. S., Semple, C. A., Devon, R. S., Clair, D. M. S., Muir, W. J., Blackwood, D. H., et al. (2000). Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human molecular genetics*, *9*(9), 1415–1423 (cit. on p. 40).

Mills, B. R. (2022). MetBrewer: Color palettes inspired by works at the metropolitan museum of art. (Cit. on p. 85).

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M., & Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic acids research*, *9*(22), 6047–6068 (cit. on p. 8).

Nash, A. (2018). *The evolutionary dynamics of genomic regulatory blocks in metazoan genomes* (Thesis). (Cit. on p. 13).

Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., & Lekschas, F. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, *593*(7858), 238–243 (cit. on p. 16).

National Centre for Biotechnology Information. (2023). *dbSNP*. Retrieved March 10, 2023, from https://www.ncbi.nlm.nih.gov/snp/. (Cit. on p. 19)

National Institute for Health and Care Excellence. (2023). *Cardiovascular disease: risk assessment and reduction, including lipid modification*. Retrieved March 10, 2023, from https://www.nice.org.uk/guidance/cg181. (Cit. on p. 71)

Navratilova, P., & Becker, T. S. (2009). Genomic regulatory blocks in vertebrates and implications in human disease. *Briefings in Functional Genomics and Proteomics*, *8*(4), 333–342 (cit. on p. 12).

Navratilova, P., Fredman, D., Hawkins, T. A., Turner, K., Lenhard, B., & Becker, T. S. (2009). Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Developmental biology*, *327*(2), 526–540 (cit. on pp. 12, 14).

NIHR BioResource–Rare Diseases Consortium. (2023). *NIHR Rare Diseases BioResource*. Retrieved March 10, 2023, from https://bioresource.nihr.ac.uk/using-our-bioresource/our-cohorts/rare-diseases-bioresource/

Nobrega, M. A., Ovcharenko, I., Afzal, V., & Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science*, *302*(5644), 413–413 (cit. on p. 12).

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–385 (cit. on p. 4).

Ohno, S. (1972). So much 'junk' DNA in our genome. *Brookhaven Symposium in Biology*, *23*, 366–370 (cit. on p. 1).

Onwordi, E. C., Halff, E. F., Whitehurst, T., Mansur, A., Cotel, M.-C., Wells, L., Creeney, H., Bonsall, D., Rogdaki, M., Shatalina, E., et al. (2020). Synaptic density marker SV2A is reduced in schizophrenia patients and unaffected by antipsychotics in rats. *Nature communications*, *11*(1), 246 (cit. on pp. 37, 48, 138).

Osimo, E. F., Cardinal, R. N., Jones, P. B., & Khandaker, G. M. (2018). Prevalence and correlates of low-grade systemic inflammation in adult psychiatric inpatients: An electronic health record-based study. *Psychoneuroendocrinology*, *91*, 226–234 (cit. on p. 35).

Osimo, E. F., Perry, B. I., Mallikarjun, P., Pritchard, M., Lewis, J., Katunda, A., Murray, G. K., Perez, J., Jones, P. B., Cardinal, R. N., et al. (2023). Predicting treatment resistance from first-episode psychosis using routinely collected clinical information. *Nature Mental Health*, *1*(1), 25–35 (cit. on p. 73).

Osimo, E. F., Stochl, J., Zammit, S., Lewis, G., Jones, P. B., & Khandaker, G. M. (2020a). Longitudinal population subgroups of CRP and risk of depression in the alspac birth cohort. *Comprehensive psychiatry*, *96*, 152143 (cit. on p. 35).

Osimo, E. F., Baxter, L., Stochl, J., Perry, B. I., Metcalf, S. A., Kunutsor, S. K., Laukkanen, J. A., Wium-Andersen, M. K., Jones, P. B., & Khandaker, G. M. (2021a). Longitudinal association between CRP levels and risk of psychosis: A meta-analysis of population-based cohort studies. *npj Schizophrenia*, *7*(1), 31 (cit. on p. 35).

Osimo, E. F., Beck, K., Reis Marques, T., & Howes, O. D. (2019). Synaptic loss in schizophrenia: A meta-analysis and systematic review of synaptic protein and mrna measures. *Molecular psychiatry*, *24*(4), 549–561 (cit. on pp. 37, 40, 48, 138).

Osimo, E. F., Brugger, S. P., de Marvao, A., Pillinger, T., Whitehurst, T., Statton, B., Quinlan, M., Berry, A., Cook, S. A., O'Regan, D. P., & Howes, O. D. (2020b). Cardiac structure and function in schizophrenia: A cardiac mr imaging study. *British Journal of Psychiatry*, *217*(2), 450–7 (cit. on p. 35).

Osimo, E. F., Perry, B. I., Cardinal, R. N., Lynall, M.-E., Lewis, J., Kudchadkar, A., Murray, G. K., Perez, J., Jones, P. B., & Khandaker, G. M. (2021b). Inflammatory and cardiometabolic markers at presentation with first episode psychosis and long-term clinical outcomes: A longitudinal study using electronic health records. *Brain Behaviour and Immunity*, *91*, 117–127 (cit. on pp. 35, 37).

Osimo, E. F., Sweeney, M., de Marvao, A., Berry, A., Statton, B., Perry, B. I., Pillinger, T., Whitehurst, T., Cook, S. A., O'Regan, D. P., Thomas, E. L., & Howes, O. D. (2021c). Adipose tissue dysfunction, inflammation, and insulin resistance: Alternative pathways to cardiac remodelling in schizophrenia. a multimodal, case–control study. *Translational Psychiatry*, *11*(1), 614 (cit. on p. 35).

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691), 239–243 (cit. on p. 8).

Palmer, D. S., Zhou, W., Abbott, L., Wigdor, E. M., Baya, N., Churchhouse, C., Seed, C., Poterba, T., King, D., Kanai, M., et al. (2023). Analysis of genetic dominance in the UK Biobank. *Science*, *379*(6639), 1341–1348 (cit. on pp. 148, 150).

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., & Lewis, K. D. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, *444*(7118), 499–502 (cit. on pp. 11, 14).

Perry, B. I., Osimo, E. F., & Khandaker, G. M. (2021a). Risk prediction in psychosis: Progress made and challenges ahead. *Biological Psychiatry*, *90*(9), 590–592 (cit. on p. 71).

Perry, B. I., Osimo, E. F., Upthegrove, R., Mallikarjun, P. K., Yorke, J., Stochl, J., Perez, J., Zammit, S., Howes, O., Jones, P. B., et al. (2021b). Development and external validation of the psychosis metabolic risk calculator (PsyMetRiC): A cardiometabolic risk prediction algorithm for young people with psychosis. *The Lancet Psychiatry*, *8*(7), 589–598 (cit. on p. 35).

Perry, B. I., Vandenberghe, F., Garrido-Torres, N., Osimo, E. F., Piras, M., Vazquez-Bourgon, J., Upthegrove, R., Grosu, C., De La Foz, V. O.-G., Jones, P. B., et al. (2022). The psychosis metabolic risk calculator (PsyMetRiC) for young people with psychosis: International external validation and site-specific recalibration in two independent european samples. *The Lancet Regional Health-Europe*, *22*, 100493 (cit. on p. 37).

Pillinger, T., D'ambrosio, E., McCutcheon, R., & Howes, O. D. (2019a). Is psychosis a multisystem disorder? a meta-review of central nervous system, immune, cardiometabolic, and endocrine alterations in first-episode psychosis and perspective on potential models. *Molecular psychiatry*, *24*(6), 776–794 (cit. on pp. 37, 149).

Pillinger, T., Osimo, E. F., de Marvao, A., Berry, M. A., Whitehurst, T., Statton, B., Quinlan, M., Brugger, S., Vazir, A., Cook, S. A., et al. (2019b). Cardiac structure and function in patients with schizophrenia taking antipsychotic drugs: An MRI study. *Translational psychiatry*, *9*(1), 163 (cit. on p. 37).

Pillinger, T., Osimo, E. F., Brugger, S., Mondelli, V., McCutcheon, R. A., & Howes, O. D. (2019c). A meta-analysis of immune parameters, variability, and assessment of modal distribution in psychosis and test of the immune subgroup hypothesis. *Schizophrenia bulletin*, *45*(5), 1120–1133 (cit. on p. 37).

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC research notes*, *12*(1), 1–7 (cit. on p. 4).

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, *47*(7), 702–709 (cit. on pp. 20, 32).

Polychronopoulos, D., King, J. W., Nash, A. J., Tan, G., & Lenhard, B. (2017). Conserved non-coding elements: Developmental gene regulation meets genome organization. *Nucleic acids research*, *45*(22), 12611–12624 (cit. on pp. 11, 12, 43, 49, 138).

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., & Canfield, T. K. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, *515*(7527), 402–405 (cit. on p. 6).

Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2020). LDpred2: Better, faster, stronger. *Bioinformatics*, *36*(22-23), 5424–5431 (cit. on p. 29).

Privé, F., Vilhjálmsson, B. J., Aschard, H., & Blum, M. G. (2019). Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics*, *105*(6), 1213–1221 (cit. on p. 27).

Psychiatric GWAS Consortium Coordinating Committee. (2009). Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, *166*(5), 540–556 (cit. on pp. 118, 119, 147).

Purcell, S. M., & Chang, C.-K. (2022). PLINK 2.0. (Cit. on pp. 85, 120).

R Core Team. (2023). *R: A language and environment for statistical computing [software]*. Computer Program. R Foundation for Statistical Computing. (Cit. on pp. 56, 85).

Radua, J., Ramella-Cravaro, V., Ioannidis, J. P., Reichenberg, A., Phiphopthatsanee, N., Amir, T., Yenn Thoo, H., Oliver, D., Davies, C., Morgan, C., et al. (2018). What causes psychosis? an umbrella review of risk and protective factors. *World psychiatry*, *17*(1), 49–66 (cit. on p. 37).

Ragvin, A., Moro, E., Fredman, D., Navratilova, P., Drivenes, O., Engström, P. G., Alonso, M. E., Mustienes, E. d. l. C., Skarmeta, J. L. G., Tavares, M. J., et al. (2010). Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proceedings of the National Academy of Sciences*, *107*(2), 775–780 (cit. on p. 12).

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., & Lander, E. S. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680 (cit. on pp. 4, 51).

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., & Ward, R. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*(6834), 199–204 (cit. on p. 25).

Rivera, C. M., & Ren, B. (2013). Mapping human epigenomes. *Cell*, *155*(1), 39–55 (cit. on p. 9).

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., Stahl, E. A., Georgakopoulos, A., Ruderfer, D. M., Charney, A., et al. (2014). A role for noncoding variation in schizophrenia. *Cell reports*, *9*(4), 1417–1429 (cit. on pp. 44, 139).

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. *PLoS medicine*, *2*(5), e141 (cit. on pp. 35, 82).

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., & Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC genomics*, *5*(1), 1–9 (cit. on pp. 11, 12, 138).

Sanges, R., Hadzhiev, Y., Gueroult-Bellone, M., Roure, A., Ferg, M., Meola, N., Amore, G., Basu, S., Brown, E. R., De Simone, M., et al. (2013). Highly conserved elements dis-

covered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. *Nucleic Acids Research*, *41*(6), 3600–3618 (cit. on p. 12).

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427 (cit. on pp. 135, 137).

Schoenfelder, S., & Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, *20*(8), 437–455 (cit. on pp. 8, 10).

Sekar, A., Bialas, A. R., De Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183 (cit. on p. 40).

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, *148*(3), 458–472 (cit. on p. 4).

Shastry, B. S. (2002). SNP alleles in human disease and evolution. *Journal of human genetics*, *47*(11), 561–566 (cit. on p. 19).

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., & Arakawa, T. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, *100*(26), 15776–15781 (cit. on p. 9).

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, *604*(7906), 509–516 (cit. on p. 39).

Siva, N. (2008). 1000 genomes project. *Nature biotechnology*, *26*(3), 256–257 (cit. on p. 30).

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485 (cit. on pp. 25, 26).

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gomez-Marin, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–375 (cit. on p. 12).

Stack, S. M., Brown, D. B., & Dewey, W. (1977). Visualization of interphase chromosomes. *Journal of cell science*, *26*(1), 281–299 (cit. on p. 7).

Strachan, T., & Read, A. P. (2018). Chapter 5: Patterns of inheritance. In *Human molecular genetics*. Garland Science. (Cit. on p. 20).

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., & Landray, M. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, *12*(3), e1001779 (cit. on pp. 76, 77).

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: Evidence from a meta-analysis of twin studies. *Archives of general psychiatry*, *60*(12), 1187–1192 (cit. on pp. 31, 32, 38, 135).

Syvänen, A.-C. (2001). Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, *2*(12), 930–942 (cit. on p. 27).

Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nature Reviews Genetics*, *3*(5), 391–397 (cit. on p. 21).

Tadros, R., Francis, C., Xu, X., Vermeer, A., Harper, A. R., Huurman, R., Kelu Bisabu, K., Walsh, R., Hoorntje, E. T., & Te Rijdt, W. P. (2021). Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. *Nature genetics*, *53*(2), 128–134 (cit. on pp. 32, 41).

Tadros, R., Zheng, S. L., Grace, C., Jorda, P., Francis, C., Jurgens, S. J., Thomson, K. L., Harper, A. R., Ormondroyd, E., West, D. M., Xu, X., Theotokis, P. I., Buchan, R. J., McGurk, K. A., Mazzarotto, F., Boschi, B., Pelo, E., Lee, M., Noseda, M., … Watkins, H. (2023). Large scale genome-wide association analyses identify novel genetic loci and mech-

anisms in hypertrophic cardiomyopathy. *medRxiv* (cit. on pp. 32, 41, 43, 50, 55, 58, 67, 68, 74, 76, 77, 102, 104, 106, 109, 110, 112, 113, 115, 137, 149, 166–168).

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, *29*(3), 288–299 (cit. on pp. 2, 3).

Tan, G. (2017). *Computational genomics of regulatory elements and regulatory territories* (Thesis). Imperial College London. (Cit. on p. 18).

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74 (cit. on pp. 2, 9).

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678 (cit. on p. 134).

Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J., & Bickmore, W. A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, *346*(6214), 1238–1242 (cit. on p. 4).

Trubetskoy, V., Pardinas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C. Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … Gershon, E. S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508 (cit. on pp. 31, 32, 37, 39, 50, 55, 58, 63, 64, 73–75, 77, 88, 118, 126–128, 134, 135, 137).

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59 (cit. on p. 147).

van de Bunt, M., Cortes, A., Consortium, I., Brown, M. A., Morris, A. P., & McCarthy, M. I. (2015). Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS genetics*, *11*(9), e1005535 (cit. on p. 24).

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., & Pennacchio, L. A. (2008). Ultraconservation identifies

a small subset of extremely constrained developmental enhancers. *Nature genetics*, *40*(2), 158–160 (cit. on p. 11).

Visscher, P. M., Yengo, L., Cox, N. J., & Wray, N. R. (2021). Discovery and implications of polygenicity of common diseases. *Science*, *373*(6562), 1468–1473 (cit. on pp. 49, 139).

Walsh, R., Buchan, R., Wilk, A., John, S., Felkin, L. E., Thomson, K. L., Chiaw, T. H., Loong, C. C. W., Pua, C. J., Raphael, C., Prasad, S., Barton, P., Funke, B., Watkins, H., Ware, J. S., & Cook, S. A. (2017). Defining the genetic architecture of hypertrophic cardiomyopathy: Re-evaluating the role of non-sarcomeric genes. *European heart journal*, *38*(46), 3461–3468 (cit. on p. 41).

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C., Clarke, D., Gu, M., Emani, P., & Yang, Y. T. (2018a). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, *362*(6420), eaat8464 (cit. on p. 52).

Wang, H.-Y., Liu, Y., Yan, J.-W., Hu, X.-L., Zhu, D.-M., Xu, X.-T., & Li, X.-S. (2018b). Gene polymorphisms of DISC1 is associated with schizophrenia: Evidence from a meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *81*, 64–73 (cit. on p. 40).

Weinberger, D. R. (2017). Future of days past: Neurodevelopment and schizophrenia. *Schizophrenia bulletin*, *43*(6), 1164–1168 (cit. on p. 37).

Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W. J., Khera, A. V., Okada, Y., Martin, A. R., Finucane, H. K., et al. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics*, *54*(4), 450–458 (cit. on p. 144).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the tidyverse. *Journal of open source software*, *4*(43), 1686 (cit. on pp. 56, 85).

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding se-

quences are associated with vertebrate development. *PLoS biology*, *3*(1), e7 (cit. on pp. 11, 12).

Yaffe, D., Nudel, U., Mayer, Y., & Neuman, S. (1985). Highly conserved sequences in the 3′ untranslated region of mrnas coding for homologous proteins in distantly related species. *Nucleic acids research*, *13*(10), 3723–3737 (cit. on p. 11).

Zeng, J., Xue, A., Jiang, L., Lloyd-Jones, L. R., Wu, Y., Wang, H., Zheng, Z., Yengo, L., Kemper, K. E., Goddard, M. E., et al. (2021). Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nature Communications*, *12*(1), 1164 (cit. on p. 32).

Zhang, Y., Li, T., Preissl, S., Amaral, M. L., Grinstein, J. D., Farah, E. N., Destici, E., Qiu, Y., Hu, R., & Lee, A. Y. (2019). Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature genetics*, *51*(9), 1380–1388 (cit. on pp. 51, 140).

Zheng, G., Freidlin, B., Li, Z., & Gastwirth, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *45*(3), 335–348 (cit. on p. 24).